

One Way or Another
Essays on Behavioural Heterogeneity and Change



Adam Ferris

Supervisors:

Professor Cameron Hepburn
Dr Stefania Innocenti

St Cross College
University of Oxford

Trinity Term 2026

*Thesis submitted as part of the requirements for the degree of
Doctor of Philosophy*

Table of Contents

Abstract	v
Acknowledgements	viii
1 Introduction	1
2 Literature Review	13
2.1 Heterogeneity in Behavioural Policy	15
2.1.1 Distributional effects	16
2.1.2 Persistence and dynamic effects	17
2.1.3 Mechanisms and channels	18
2.1.4 Ethics of Behavioural Policy	18
2.1.5 Generalisability and Scalability	22
2.2 Methods for Behavioural Policy	27
2.2.1 Study Environments	27
2.2.2 Replicability	29
2.2.3 Identification Strategies	32
3 Paper 1	37
<i>Should I Stay or Should I Go: Intentions to Freelance Across 16</i> <i>Countries</i>	37
3.1 Paper 1 (Front Matter)	38
3.2 Introduction	39
3.3 Survey Design	43
3.4 Data	48
3.4.1 Sample Selection	48
3.4.2 Data Description	49
3.5 Determinants of Intentions to Freelance	53

3.5.1	Empirical Specification	53
3.5.2	Main Results	55
3.6	Robustness	59
3.7	Heterogeneity	62
3.8	Discussion and Conclusion	67
3.9	Appendix: Additional Descriptive Evidence	71
3.10	Appendix: Additional Robustness	79
3.11	Appendix: Survey Questions	86
3.12	References	90
4	Paper 2	98
	<i>Do Food Subsidies Change Diets for Good? Evidence from Subsidy</i>	
	<i>Introduction and Withdrawal</i>	98
4.1	Paper 2 (Front Matter)	100
4.2	Introduction	101
4.3	Methods and Data	105
4.3.1	Experimental Design	105
4.3.2	Intervention	107
4.3.3	Data	108
4.3.4	Dependent and Other Variables	109
4.3.5	Research Questions	110
4.3.6	Estimation Strategy	111
4.4	Results	114
4.4.1	Descriptive Statistics	114
4.4.2	Adding a Vegetarian Meal Subsidy	115
4.4.3	Removing a Vegetarian Meal Subsidy	117
4.5	Heterogeneity Analysis: Effects of Underlying Diet Preferences . .	119
4.6	Event Study: Intervention Spillover	122
4.7	A Note on Cost-effectiveness of subsidy introduction	123

4.8	Discussion and Conclusion	127
4.9	References	131
4.10	Appendix: Main Results: Adding Subsidy	137
4.11	Appendix: Results Shorter Period	144
4.12	Appendix: Pre-trend of vegetarian meals	148
4.13	Appendix: General Information about meals and colleges	153
5	Paper 3	155
	<i>Identity Activation and Choice Architecture as Complements: Ex-</i> <i>perimental Evidence</i>	155
5.1	Paper 3 (Front Matter)	156
5.2	Introduction	157
5.3	Experiment	161
5.3.1	Experimental Setting and Shopping Task	161
5.3.2	Interventions	162
5.3.3	Experimental Design	164
5.3.4	Procedure	167
5.4	Data	169
5.4.1	Sample	169
5.4.2	Variables	172
5.5	Hypotheses	175
5.6	Estimation Strategy	178
5.6.1	Willingness-to-Pay Estimation Strategy	179
5.6.2	Basket Emissions Estimation Strategy	180
5.7	Results	184
5.7.1	Descriptive Statistics and Demand for Eco-Sorting	184
5.7.2	Effect of Identity Cue on Eco-Sorting WTP	187
5.7.3	Effect of Eco-Sorting on Basket Emissions	189
5.7.4	Complementarity at the Point of Choice	193

5.8	Managerial and Policy Implications	196
5.8.1	Unintended Effects on Spend, Time, and Satisfaction . . .	196
5.8.2	Distribution Effects	197
5.8.3	Ethical Intervention Design	198
5.8.4	External Validity	199
5.9	Discussion and Conclusion	201
5.10	References	204
5.11	Appendix: Extra Results	210
5.12	Appendix: Managerial and Policy Implications	226
5.13	Appendix: Food and Basket Variables	230
5.14	Appendix: Experiment Visuals	232
5.15	Appendix: Multiple Price List and Willingness-to-Pay Variable Construction	237
5.16	Appendix: Worked Examples: from MPL choices to IPW propen- sity scores	244
5.17	Appendix: Survey Questions and Coding	247
5.18	Appendix: Attention and Mechanism Checks	249
6	Discussion and Conclusion	258
6.1	Overview	258
6.2	Themes and Threads of the Three Papers	263
6.3	Limitations	270
6.4	Behavioural Policy Design and Evaluation Conclusions	272
6.5	Closing Remarks	274
	Bibliography	276

Abstract

Choice environments, the contexts in which our preferences are shaped and our decisions made, are never neutral. Whether deliberately designed or arising as market circumstances and institutional decisions, choice environments bend our behaviour one way or another. And there is also variation among individuals in how they respond to these environments. This is behavioural heterogeneity, and I use it in the sense of taking two forms here. The first is heterogeneity in the broad sense of circumstances: attention, planning horizons, and other psychological factors but also neo-classical constraints such as income, job and family situations all operating at the moment of interaction with the choice environment in question. The second is heterogeneity in priors and predispositions, this includes, but is not limited to, the beliefs, tastes and habits a person brings with them, which shape how they respond to the same environment. This thesis argues, and offers three essays showing, that these heterogeneity forms should be central to the design and evaluation of choice environments, rather than treated as an auxiliary concern.

Despite a maturing literature on behavioural change and effects of choice environments, behavioural policy research still often focusses on effects for the population as a whole rather than separately for different groups. This omission risks ignoring distributional, ethical, dynamic and generalisability concerns.

This thesis addresses this through three empirical papers that highlight the importance of studying behavioural heterogeneity and change in choice environments. They progress from *observing* how an undesigned choice environment sorts workers across labour-market arrangements, to *evaluating* a deliberate price intervention in the food-choice context, to *designing* a low-cost behavioural intervention for sustainable consumption. They also span the two behavioural heterogeneity sources of circumstance and predisposition.

The first paper highlights distributional risk arising from a background choice environment and behavioural heterogeneity in circumstance, arguing for behavioural policy to account for behavioural heterogeneity analysis and policy takeaways. Using a 16-country survey and logistic regression to study intentions to transition from formal employment into freelancing, it identifies socio-demographic, labour-market and psychological factors that jointly predict stronger transition intentions, consistent with more vulnerable cohorts being “pushed” into this precarious and less regulated job type rather than opportunity “pulling” them in.

The second paper shows that deliberate interventions can have weak average treatment effects, be unevenly distributed and non-persistent. The paper shows the effects of behavioural heterogeneity in predisposition. Exploiting a multi-year natural experiment across four university cafeterias, it evaluates the introduction and withdrawal of vegetarian meal subsidies using individual-level transaction records and pre-registered difference-in-differences and event-study designs. Effects are small and concentrated among already largely vegetarian diners, with no detectable response among habitual meat-eaters. Withdrawal effects are symmetric and immediate, indicating the subsidy operates through standard price-demand effects rather than preference formation, with limited cost-effectiveness relative to alternatives in the literature.

The third paper shows how a deliberately designed intervention can be designed to account for and evaluate behavioural heterogeneity in effects, studying general heterogeneity, as well as mechanism, ethics and scalability points. A pre-registered, incentive-compatible framed field experiment in an online supermarket combines identity-salience cues with an environmental-impact sorting layout of products. Combining the interventions raises willingness-to-pay for the sorting option, which when adopted delivers large emissions savings, with effects largest among low-search-effort shoppers. A high share of marginal adopters suggests a

cheap, autonomy-preserving and transparent “self-nudge” can generate economically meaningful environmental gains.

Taken together, the papers make three contributions. Empirically, they document research- and policy-relevant behavioural heterogeneity in how populations respond to both designed and undesigned choice environments. Methodologically, through pre-registration, channel measurement, secondary analysis and incentive-compatible elicitation, they set out elements of credible behavioural-economic evaluation addressing replication, generalisability and scalability. For policy, they argue that behavioural heterogeneity in results through general, distributional, ethical and dynamic (short- to long-term) criteria should be embedded in intervention design and revisited throughout deployment, and discuss steps for helping target populations “do the right thing willingly”.

This thesis is submitted as an integrated thesis of publishable papers.

Acknowledgements

I could not have completed my DPhil without standing on a human pyramid of many people, each lending a hand to lift me higher.

I thank my supervisors, Professor Cameron Hepburn and Dr Stefania Innocenti. Your guidance, expertise, and time throughout this whole DPhil process has been invaluable. Thank you for overseeing my work and my development as an academic. Your feedback and advice will undoubtedly remain part of my researcher-mindset going forwards.

To my co-authors Drs Andrea Byfuglien, Michael Clark, Stefania Innocenti and Paul Lohmann, it has been a pleasure and a delight working with you all. Thank you for making my research more rigorous and an enjoyable process. My work has also only been possible thanks to the intellectual generosity of Neil McGowan and Dr Ellen Michalski. Thank you for your time and financial support through the Kellanova-SSEE project and scholarship. Cheryl Song, I also owe you a debt of gratitude for your support and design of this collaboration and scholarship.

Whatever the dosage of nature and nurture that has got me this far in life, including their support and encouragement, I am extremely grateful to my parents. Mum, Dad, Philip, and Emma thank you for your continued and unwavering love, I cherish it dearly! My curiosity, determination and ambition that have lead me to the end of this DPhil are all thanks to you.

I am also especially grateful to Noel and Linda who – through an unlucky draw for them and a lucky one for me – were chosen as my godparents. To my wider family who can self-assign to our family categories, those in Canada, the US, Ireland, and the UK, thank you for letting me be part of your lives.

I have many friends that have put up with me and supported me through this DPhil journey. My college family, Cal and Lonie. My Manor Road Building

SSEE friends, Siyu, Joseph, Johanna, Fulvia, Mathias, Ebba, Mads and Swantje. And of course, the INET gang next door, Dorothy, Susan, Niki, Jonathan, Sam, Emilien. Valentina, Karolina, Ximeng, Joris, Christian, Jan, Owen, Tim, Yuke, Daniel, Fernanda. Thank you all. With particular thanks to Benjamin and Rohit.

I owe a special thank you to Mary, Alena, James, Ben, Jorn, Kit, Tom, Wojciech, and Andrew. As well as Anu, Eloïse, Claire, Federico, Helena, Marianne, Moritz, and Martina. Thank you all for the laughs, friendship and support over the years.

I would like to thank my college, St Cross College for being my home through my Masters and DPhil. Thank you to the team for taking care of me. Thank you to the Wolfson College Boat Club for keeping me afloat and away from the office for a little bit each week. And thank you for the huge community of students and friends that I have met through the boat club.

I would like to also thank several groups of mentors and institutions that have offered jobs for me to help fund my DPhil and develop further as an academic. Thank you to Dr Noah Bacine and the Centre for Experimental Social Sciences, Nuffield College Oxford. Thank you to the Saïd Business School for the education and the teaching opportunities. Thank you Dr Kazbi Soonawalla for being a wonderful mentor and supporter at SBS. I would also like to thank Dr James Carroll and Professor Eleanor Denny at Trinity College Dublin, for having me as a research assistant. Similarly, thank you to Professor Ian Goldin at the Oxford Martin School for having me as a research assistant and showing me the importance of real-world impact work.

Thank you to Oxford, the city and university for being my home for the last several years.

Chapter 1 | Introduction

Overview and Motivation This thesis argues that behavioural heterogeneity in how individuals respond to choice environments, whether based on circumstance or predisposition, should be central to the design and evaluation of choice environments, rather than treated as an auxiliary concern. In the three studies of this thesis, I show the prevailing behavioural heterogeneities and their important implications for behavioural policy design and evaluation through their consequences for distributional, ethical, dynamic and generalisability concerns.

Choice environments shape individuals' preferences and choices, for example through which options are offered, how they are priced, ordered or perceived and which information is made salient. However, these choice environments rarely affect everyone in the same way. I use the term choice environment to avoid confusion with 'choice architecture' definitions which most recently focus on Nudges (Thaler, 2016; Johnson et al., 2012; Mertens et al., 2022; Banerjee and John, 2025). I define 'choice environment' as broadly covering both *structural* features of a choice set (what is available, at what price, under what constraints) and *presentational* features (framing and messaging, ordering, salience, defaults), and can arise from deliberate design or from the incidental rules of policies, platforms, and markets. It is important to recognise that a choice environment is never neutral, it (un)knowingly favours some choices over others (Shafir, 2013).

I use the term 'behavioural heterogeneity' to refer to the differences in behavioural choices and responses across individuals when faced with the same choice environment. Two sources of this heterogeneity are considered in this thesis. The first is the heterogeneity in *circumstance*, which includes the income, family, psychological and other constraints operating at the moment an individual meets the choice environment in question. The second is the heterogeneity in *predisposition*, which encompasses the beliefs, tastes, experiences and habits a person

brings with them, shaping how they respond to an environment.

It is true that some parts of the *circumstances* are not behavioural mechanisms but traditional neo-classical ones, such as income effects. However, I adopt a broad definition of behavioural heterogeneity, as it has previously been shown that neo-classical mechanisms can interact with cognitive factors in behavioural change. Examples include financial situations narrowing attention and planning horizons (Shah et al., 2015; Barberis et al., 2006), or when income position under job scarcity affects impatience and job search efforts (DellaVigna and Paserman, 2005; DellaVigna et al., 2017).

With different empirical data sources and types, I use econometric methods to examine how a behavioural policy researcher might approach choice environment evaluation and design under these real behavioural heterogeneity constraints and show their important impacts for behavioural policy. I do so across cases of job selection (Chapter 3), vegetarian meal choice (Chapter 4), and self-nudging into selecting a different choice environment (Chapter 5).

Throughout this thesis, I consider behavioural policy research to encompass any application of behavioural science insights to the design of policies and interventions, whether for public or private institutions, and whether designed or happenstance by an individual, group, government, or market. This broad interpretation also follows those that some use for behavioural public policy and behavioural economic research (Shafir, 2013; Thaler, 2016; Weber, 2013).

A successful intervention or policy should be evaluated not only by its efficiency, but by whether its design shapes decisions in ways that enhance social welfare (Shafir, 2013; Weber, 2013; Lades and Delaney, 2022) and are accepted by society (Sunstein et al., 2019). Pursuing that requirement needs more than an average effect. An average masks who is responsive and who is not, and with it the distributional, ethical, dynamic and generalisability concerns that determine

whether a policy is doing genuine good. Behavioural heterogeneity is therefore not a robustness check to be reported after the headline estimate, but the object of evaluation itself. This is the importance of a broader design, analysis and evaluation behavioural policy.

To measure these before the policy adoption stage the behavioural researcher might run (pilot) studies, and the generalisability and scalability of these results adds a wedge of complexity. All together, the researcher's decision process is far from linear. I discuss this further in the Literature Review and Discussion (Chapters 2 and 6) of this Thesis.

Background on Behavioural Policy Research It is important to highlight that behavioural policy research does not reject traditional economics and policy instruments but often fine-tunes policy designs and expectations by incorporating human behaviour from psychology and economics. While traditional economics models assume perfect rationality, self-interest and utility-maximisation, behavioural research since Simon et al. (1972), Tversky and Kahneman (1974) and Thaler (1980) show that individuals have bounded rationality, limited attention and cognitive constraints.¹ These lead to people perhaps wishing to choose optimally, but failing because of these biases driven by internal factors and external ones (DellaVigna, 2009; Weber and Johnson, 2009) for example self-control and temptation issues (Toussaert, 2018).

Behavioural policy, in addition to traditional incentives (e.g. monetary incentives or regulating options provided), considers designs that account for biases that lead people to choose non-optimally with respect to their (long-term, underlying or intrinsic) preferences (Weber, 2013; Thaler, 2016; Johnson et al., 2012).

¹Of course, these are not the origins of behavioural understanding. Adam Smith already discussed behavioural aspects in the *Theory of Moral Sentiments* (Smith, 1759; Ashraf et al., 2005). And prior to its academic roots, the street vendors of Ancient Rome surely knew the marketable value of the right frame for a trade deal, while illusionists and street performers have always exploited our attention and other cognitive limits to dazzle and bemuse.

This is the role of behavioural policy makers, being sensitive to the benchmark behavioural impacts of ‘non-neutral’ decision environments and contexts (Shafir, 2013; Thaler, 2016; Johnson et al., 2012; Laibson and List, 2015). Furthermore, not only being aware that policy can bend preferences through framing, regulation of options available and information provided, and heightening salience of an option’s or person’s attributes, but that these effects will differ across individuals should be the crux of behavioural policy research.

It is also worth noting, by designing for equitable and socially beneficial outcomes, the policy makers should avoid ‘sludge’ for the whole or a specified target (vulnerable) population. Sludge refers to interventions that (by choice environment design or happenstance) damage social welfare (Thaler, 2018; Weber, 2013; Lades and Delaney, 2022; Thunström et al., 2018).

Choice environment and the three essays The non-neutrality and behavioural heterogeneity of choice environments is visible across domains (Mertens et al., 2022; Shafir, 2013; Thaler and Sunstein, 2021; Sunstein, 2014). This thesis focusses on two contrasting domains, labour and food choices. These illustrate the pervasiveness of choice environments across both high stakes often infrequent, and then low stakes frequent decision making.²

For labour, whether considering the choice of a stable job, a gig or more frequently the amount of effort to provide at work, these are typically higher stake decisions. Monetary incentives are a traditional economic lever discussed in this domain, but their effects (both directionally and in magnitudes) depend on the context in which they are applied. Bénabou and Tirole (2006) and Gneezy et al. (2011) describe monetary incentives ‘crowding out’ intrinsic incentives. Empirical work shows that incentives can work in non-routine tasks, but could counter

²Albeit the low stakes food decisions for the individual have important high stakes at the aggregate level, i.e. as critical behaviour change for climate mitigation (Clark et al., 2020).

creative exploration and intellectual output in jobs with intrinsically motivated (self-selected) employees (Englmaier et al., 2017). Freelance platform-based ranking and priority systems, indirectly incentivise gig workers to overwork and lose work-life balances sought after when choosing freelancing (MacDonald and Giazitzoglou, 2019; Wood et al., 2019). Framing also matters. Hossain and List (2012) demonstrate that the salience of an incentive is stronger when its contract design is framed as a penalty rather than a gain, in line with prospect theory.

Framing for fairness also moderates labour effort (Fehr et al., 2009; Cohn et al., 2015). Contexts are important to consider, financial scarcity can bring narrow focus to shorter-term benefits and planning horizons (Shah et al., 2015; Mullainathan and Shafir, 2013). Heterogeneity in income situation and impatience shapes job search effort (DellaVigna and Paserman, 2005; DellaVigna et al., 2017), and underscores how policy design interacts with behavioural biases.

Similar mechanisms apply in lower-stakes frequent domains such as food choices. The An (2013) meta-analysis shows that most subsidies on healthy foods increase their uptake, but highlights difficulties in inferring patterns or heterogeneities across the existing studies due to varying timelines, specific samples used and availability of substitutes. In environmental food policy, Pizzo et al. (2024) show that subsidy vouchers outperform informational nudges in food markets. Garnett et al. (2021) finds that price interventions in cafeterias mainly affect uptake among those already consuming vegetarian meals, and only with a small effect size. Perino et al. (2014) discuss how framing a discount as a government-led discount, rather than a neutral-framed discount, reduces effectiveness, perhaps by crowding out motivation. Salient signposting of incentives (Zizzo et al., 2021; Chetty et al., 2009), presentation (grouping or ordering) of items (Panzone et al., 2021a; Pechey et al., 2022a) and more widely whether people agree with the overarching objective of a policy to begin with matters (Pechey et al., 2022b; Funke

et al., 2022) for effectiveness of incentives. Although food choices are low stakes, for the individual, from an economic cost perspective, they are culturally charged and can be difficult to nudge with informational interventions e.g. with social norms (Pollicino et al., 2025). But, demand-side changes in the food system, matter and, crucially, are needed to achieve net zero (Clark et al., 2020, 2022).

Behavioural Policy and the essays’ methodologies From a methodological perspective for this thesis, the past two decades have produced a growing body of work designing and testing behavioural intervention effects, with some critiques such as whether their claimed effects sizes are credible (magnitude or M-size errors) (DellaVigna and Linos, 2022; Andrews and Kasy, 2019; Mertens et al., 2022). I discuss these further in the Literature Review and Discussion Chapters (2 and 6) of this Thesis.

Another concern is that success should not be measured solely by average treatment effects. Decisions are context dependent, shaped by internal and external factors (Weber and Johnson, 2009). Consequently, policies may operate differently across populations or settings (Thunström et al., 2018). Advancing the field requires understanding for whom and under what (direct and indirect) conditions these effects emerge, and subsequently managing policies to be both effective, well-targeted and equitable (Weber, 2013; Gneezy et al., 2011; Schimmelpfennig et al., 2021). Generalisability and scalability of results, affected by behavioural heterogeneity, is a key discussion point for policy makers. Notably when behavioural early pilot studies are based on small specific population samples, and only study the effects of one side of the demand-supply side equation. Other aspects to consider from the behavioural heterogeneity point are the mechanism and distributional effects at play in these policies. These are discussed in the Literature Review and Discussion Chapters of this Thesis.

Contribution of this Thesis This Thesis presents three papers chosen to highlight the importance and consequences of behavioural heterogeneity for behavioural policy evaluation and design. Its contribution follows three paths. Empirically, the papers present behavioural heterogeneity in policy relevant choice environments, whether deliberately designed or market circumstances. Methodologically, I present several elements of rigorous and credible behavioural policy and behavioural heterogeneity analysis and evaluation, such as pre-registration, mechanism assessment, secondary analysis and the use of incentive-compatible data results. Finally, for policy, the papers argue that distributional, ethical, dynamic and generalisability consequences of behavioural heterogeneity should be central to policy design, analysis and evaluation, rather than secondary after studying a main, average, effect.

Below, I set out the overviews and methods of the papers in this integrated thesis, and the novel questions each asks that tie back to this framing.

Paper Overviews: Across all three papers, the impact of the choice environment design on the effectiveness of the policies is underscored. Behavioural heterogeneity from the policies are highlighted for their important and careful consideration for discussions on effective and equitable behavioural policy. Together, the three papers trace an arc from the *incidental* choice environments embedded in market and platform rules (Paper 1), through a *deliberate* but imperfectly distributed and non-persistent policy intervention (Paper 2), to a *deliberately designed* intervention evaluated against the full set of considerations (Paper 3).

Paper 1 (Chapter 3): Paper 1 engages the thesis theme by demonstrating that the incidental choice environment of the labour market creates distributional risk, making the case for policy evaluation that accounts for behavioural heterogeneity in effects. The paper investigates the results of the design and current regulation of freelancing work and opportunity access (e.g. through digital platforms)

shaping individuals' labour market choices, underscoring that such environments are not neutral. Using a large, novel, and international (16 country) dataset of survey answers from employed workers, the paper describes the behavioural, socio-demographic and occupational factors associated with intentions to transition from formal employment to freelancing. The work incorporates novel metrics such as task-type and cognitive reflection to contribute to research in modern self-employment research. The work highlights behavioural heterogeneity in freelancing interest, with economically and situationally vulnerable groups disproportionately more likely to state intention to freelance within the next 12 months. This suggests that platform access and regulation are a form of choice environment with potential welfare and inequality implications.

Paper 2 (Chapter 4): Paper 2 engages the thesis theme by showing that even deliberate interventions can have weak average treatment effects, be unevenly distributed in their effects and non-persistent. The paper examines how pricing policies in real-world food settings shape consumers' dietary choices. Using individual-level sales data from a natural-field cafeteria environment spanning multiple years, the paper examines how both an introduction and a withdrawal of subsidies for vegetarian meals affects purchasing behaviour across individuals with differing baseline dietary preferences. The multiple years of data enables a novel contribution of policy persistence to the food literature, for both pricing policy changes, the withdrawal of a subsidy in particular being rarely studied. The paper shows that while subsidies modestly increase overall uptake of vegetarian meals, these effects are concentrated among consumers who were already predominantly vegetarian, with no change in behaviour observed among the habitual meat-eaters. When subsidies are withdrawn, prior high-vegetarian meal consumers significantly reduce their vegetarian purchases. All of these changes are stable following the policy shift. These heterogeneities raise questions about the efficiency and skewed benefits of pricing policies, and their limited persistence,

in the way they are currently designed.

Paper 3 (Chapter 5) Paper 3 engages the thesis theme by deliberately designing an intervention with behavioural heterogeneity, persistence, mechanism, ethics and scalability considerations central to its focus. The paper explores how a small informational cue, can lead to self-nudging into a change of choice environment and large outcome changes, and highlighting the non-neutral and potential tipping point-style effects of choice environment. Using an incentive-compatible design the paper studies the heterogeneous responses to a commercial behavioural policy that re-sorts food products within their categories according to environmental (eco) impact, changing the order and salience of products to be considered. The results show that although all consumers' groceries' emissions are reduced across several factors, search effort is a key heterogeneity moderator. The paper further investigates the willingness-to-pay for the eco-sorting and shows that an extra identity cue message can increase the interest in that sorting option. This suggests that a small message change to the decision environment can amplify acceptability and scalability of adopting eco-ranking policies.

Methods Although all three papers are empirical in nature, to answer the different research questions set out for each paper, data sourcing, research flow and empirical methods diverged.

Data sourcing: To study the variation in individuals' intentions to engage in freelance work, Paper 1 uses a large, international survey dataset pre-dating my DPhil as part of a department collaboration with Zurich Insurance. This novel dataset captures a large batch of socio-demographic factors, psychological traits, work situational factors and more. These survey questions were all asked to participants across 16 different countries. Previous analyses using the same data include Innocenti and Golin (2022) and Innocenti et al. (2023). Paper 2 exploits a natural field experiment (Harrison and List, 2004; List, 2011) occur-

ring at the University of Oxford for its data source. University cafeteria sales data have previously been used in behavioural economic research (Garnett et al., 2019, 2021; Schulze-Tilling, 2025; Lohmann et al., 2022). We collect anonymised individual-level meal sales data from four colleges to examine how subsidies influence food choices. Following Harrison and List (2004), this setting benefits from observing real-world behaviour in a natural environment, without a risk of any experimenter-demand effect. Because students cannot switch between college cafeterias, our design also avoids displacement effects, a concern in policy evaluation (Lacetera et al., 2012). However, external validity is limited by the student context and cafeteria prices already being unrealistically low (Harrison and List, 2004). In Paper 3, I gain the most experimental control through an academic collaboration that provided me with access to a naturalistic research supermarket environment (“Woods”), and a commercial collaboration awarding me research funding for participants and choice incentives. For this paper, I designed a panel data experiment, collecting observations from participants across two shopping events and randomly assigning treatment groups at the second event. Although participants are aware of the study, their decisions are incentive-compatible, and the data are best described as arising from a quasi-framed field experiment (List, 2011), offering higher ecological validity than a standard laboratory experiment.

While the literature often debates which data source is best suited to particular research questions, in practice the decision can often be influenced by administrative issues (data protection or access) and budgetary constraints.

Research Integrity: For all three Papers, additional methodological steps were taken to ensure research integrity. All studies underwent review by the University of Oxford Central University Research Ethics Committee (CUREC), which assessed the ethical rigour and data security standards of research involving human participants, in line with the University’s policy. A second measure was the

pre-registration of research plans, which were time stamped and submitted to an online repository, for papers 2 and 3, prior to data collection. Pre-registration is increasingly encouraged in the social sciences. The movement aims to enhance transparency and credibility by requiring researchers to pre-specify their research questions, hypotheses, and analysis strategies (Imai et al., 2025; Logg and Dorison, 2021). Many leading journals now require pre-registration of experimental work, if the authors wish their work to be considered for publication at their journal. Pre-registered work is also linked to greater replicability of research and less researcher bias, although bias in results driven by the researchers' declared analysis strategy will still remain (Huntington-Klein et al., 2021). To further increase transparency, I will also provide the data used for the analysis in the three papers with their respective publications.

Statistical Methods Different estimation and inference strategies are required across the three papers.

Paper 1 uses cross-sectional logistic regressions for the international dataset, as the dependent variable (intentions to engage in freelancing) is coded as binary. We include evidence-based controls for socio-demographic, psychological, and work-context characteristics that may independently influence this outcome. We also included country-level and occupation-level dummies to absorb the correlated impact of either of these on our main variables of interest.

Paper 2 uses panel data difference-in-difference. Even though the dependent variable is binary, linear estimation is preferred for DiD settings that include fixed effects (time and individual, in our case) and clustered errors (Angrist and Pischke, 2009; Bertrand et al., 2004). Because the pricing-policy treatment is assigned at the cafeteria level, standard practice is to cluster errors at that level to account for within-cafeteria correlation (Cameron and Miller, 2015; Abadie et al., 2023). However, our estimation strategy is based on only one treatment cafeteria

paired with two control cafeterias, making both wild-cluster bootstrapping and sub-clustering invalid for inference (MacKinnon and Webb, 2018; Roodman et al., 2019; Cameron et al., 2008; Cameron and Miller, 2015). Following precedence in the food-choice literature we cluster at the individual level to account for non-independence in individual purchasing behaviour (Lohmann et al., 2022; Garnett et al., 2019), and accounting for strong serial correlation of individuals' choices.³ Ultimately, we report all inference with the appropriate caveats and transparency. Additionally, the paper includes event-study analysis to visually investigate dynamic effects from the pricing policy changes. The paper stays cautious about causal and generalisable claims in this quasi-experiment.

Paper 3 applies both a two-period DiD with person-level fixed effects and a cross-sectional study. Treatment is randomly assigned at the individual level (for the identity cue and quasi-random with Inverse Probability Weighting for the sorting treatment). The dependent variable is the environmental score of each participant's shopping basket across two weeks. Because these basket scores have a long right-hand tail, we estimate a log-linear specification as done in previous and similar work (Panzone et al., 2021b,a). For inference on 2-period DiD where treatment is at the individual level, heteroscedasticity robust errors are appropriate.

The remainder of the Chapters of this Thesis include the Literature Review of the thesis (Chapter 2), followed by the respective papers of this Integrated Thesis (Chapters 3, 4 and 5), and finally the Discussion and Conclusion (Chapter 6) ends the Thesis.

³While traditional clustering views (Bertrand et al., 2004; Abadie et al., 2023) have argued to cluster at the treatment or coarsest level, Roth et al. (2023) have argued that clustering at a finer level than the treatment assignment level can be warranted and a more useful choice.

Chapter 2 | Literature Review

This thesis states that individuals respond differently to the same choice environments, and that this behavioural heterogeneity should be central to the design, analysis and evaluation of behavioural policy, rather than an auxiliary concern reported after the study of an average effect, or ignored altogether. Behavioural heterogeneity leads to several important constraints for behavioural policy, such as distributional consequences, persistence of effects, ethical considerations, and generalisability and scalability challenges. This chapter reviews the literature on these constraints through the two sources of behavioural heterogeneity, and defends their significance for the contribution of the thesis essays that show their impact in three independent papers. The literature review also serves to introduce the methods used in the papers, and to give an overview of methods available to behavioural researchers concerned with these constraints.

Behavioural policy design and analysis at its core is about understanding how human behaviour shapes our actions and reactions to policies, beyond the standard economic and policy models of rational agents. With these insights, we can design policies that are more effective, efficient, and equitable, but we need to be wary of potential threats to the quality of our results, and of the fact that these rarely extrapolate evenly across a population.

As defined in the Introduction of this Thesis, I broadly distinguish between two sources of behavioural heterogeneity. Heterogeneity in *circumstance* refers to the income, family, attention and other constraints operating at the moment of interaction with the choice environment. Heterogeneity in *predisposition* refers to the beliefs, tastes, experiences and habits a person brings with them. Some psychological factors could appear in both of these sources. Impatience and impulsivity, for instance, are in part stable traits to which a person is predisposed, but they are also induced by circumstance, with financial scarcity narrowing

attention, search effort and planning horizons at the moment of choice (Shah et al., 2015; Mullainathan and Shafir, 2013; DellaVigna and Paserman, 2005). Both sources need to be considered for good behavioural policy evaluation and design.

To frame the literature review on behavioural policy below, it is important to recall that I use the broad definition of ‘choice environment’, as set out in the Introduction of this Thesis. The term covers not only deliberately designed behavioural interventions, but also the incidental choice environment embedded in the rules of policies, platforms, and markets. The latter is an important inclusion, as discussed with Paper 1 (Chapter 3), which shows that even the incidental choice environment of the labour market can shape individuals’ labour market choices and have consequences on wider labour market and societal outcomes. This underscores that such environments are not neutral and have differing effects across individuals, so they are worth evaluating through a behavioural lens that asks not only whether behaviour changed on average, but for whom it changed and for whom it did not. The same takeaway is drawn for the more deliberately designed interventions presented in Paper 2 (Chapter 4) and Paper 3 (Chapter 5).

Consequently, the evaluation and design considerations developed in this chapter, namely the distributional, dynamic, ethical and generalisability consequences of behavioural heterogeneity, apply with equal force across this thesis, whether the choice environment is deliberately engineered or inherited from existing institutions and markets.

The Literature Review Chapter is organised in two main sections. Section 2.1 discusses behavioural heterogeneity and the evidence measures which I argue must be part of policy evaluation and analysis. These include the distributional, dynamic, ethical and generalisability consequences of that heterogeneity. Section 2.2 presents the methodological toolkit available to the behavioural policy researcher,

and as used in the papers in this thesis.

2.1 Heterogeneity in Behavioural Policy

Behavioural policy seeks to understand and implement interventions from deliberate or incidental choice environments from public or private institutions. Behavioural policy design and analysis at its core is about understanding how human behaviour shapes our actions and reactions to policies, beyond the standard economic and policy models of rational agents. For the purposes of this thesis, I consider the effects of behavioural policy as acting along two pathways. The first is the *circumstance* in which an individual finds themselves when exposed to a choice environment, that is, the state they are in. The second is the *predisposition*, which also affects their response but through the individual's inherent characteristics and tendencies (e.g. priors, taste, experiences, cognitive biases). These pathways have been previously identified as crucial for understanding behavioural responses with distributional impacts on ethics and fairness (Lades and Delaney, 2022), the generalisability of policy studies and for designing effective interventions (List, 2020; Al-Ubaydli et al., 2017; DellaVigna and Linos, 2022).

Through circumstances, behavioural heterogeneity occurs at the moment of interaction with the choice environment from aspects of income and budget, family requirements, but also cognitive aspects such as attentional capacity or other traits cued and amplified by the choice environment (Simon et al., 1972; Shah et al., 2015; Mullainathan and Shafir, 2013; DellaVigna and Paserman, 2005). For example, the same food tax policy may have a greater relative financial burden on some, or not be salient for all, causing distributional effects (Funke et al., 2022; Klenert et al., 2023; Chetty et al., 2009). Predisposition, I use as a broad term to cover prior beliefs as well as tastes, habits and experiences. A choice environ-

ment can cue a previous memory of an experience and affect choice preference (Bordalo et al., 2020; Malmendier and Nagel, 2011). An information intervention that changes our prior of an attribute can change our behaviour (Haaland et al., 2023; Epperson and Gerster, 2026). The subsections below present the literature around both sources through their distributional, dynamic, ethical and generalisability consequences.

2.1.1 Distributional effects

Even when an average effect measurement is informative about the effectiveness of a policy, with a representative policy population, it can mask underlying biases that favour or hurt different groups. I later discuss this concern through the FORGOOD framework (Lades and Delaney, 2022) in Section 2.1.4, and it is also visible in work on the regressive effects of food carbon taxes on lower-income households (Funke et al., 2022; García-Muros et al., 2017; Klenert et al., 2023; Säll, 2018). To account for the heterogeneity of effects in the ATE, methods like the synthetic difference-in-differences (Arkhangelsky et al., 2021) exist, but this does not inform the policy researcher of the behavioural heterogeneity aspects. A methodological response instead is to estimate the Conditional Average Treatment Effect (CATE) across relevant subgroups of the policy population. For example, Epperson and Gerster (2026) measure the CATE of a messaging nudge between those who wished to avoid it and those who sought it out. D’Acunto et al. (2021) separately use interaction effects and a sample split to study the effects of inflation expectations based on gender and whether the individual is a primary grocery shopper, and so exposed to the frequent inflation effects on grocery goods.

I conduct a similar analysis as Epperson and Gerster (2026) with my experiment and random assignment in Paper 3 (Chapter 5). The same point also bears on the choice of estimator. Difference-in-differences with unit fixed effects (FE) identifies the treatment effect from within-unit variation around the policy change,

absorbing out any time-invariant differences across individuals or groups. This is precisely what makes it useful for causal identification, but it is also the heterogeneous information that we want to know about in behavioural policy, for example heterogeneity in income, education, or baseline preference groups. To recover this heterogeneity we can interact the treatment with the subgroup factor of interest (to study how it moderates the treatment effect), or estimate the CATE directly. Both Paper 2 (Chapter 4) and Paper 3 (Chapter 5) apply these strategies.

2.1.2 Persistence and dynamic effects

Another important measurement point is the time frame over which an intervention is evaluated. Behavioural intervention effects, notably informational nudges, have been shown to fade over time in some cases (Allcott and Rogers, 2014; Epperson and Gerster, 2026), or to persist when messaging is repeated or where the nudge encourages household investment in more efficient technologies (Bernedo et al., 2014; Ferraro et al., 2011). One explanation is that the salience of the change fades as it is ‘re-weighted’ in our memories (Bordalo et al., 2020). Persistence is a key part of the cost-effectiveness of behavioural policies (see Section 2.1.5), and it is important to measure the persistence of effects over time, rather than just the short-term effect. It is also central to policy design, where the aim is often persistence during the policy but also beyond it. This forms part of the rationale for the policy design in Paper 3 (Chapter 5), which implements the use of a default-like intervention which is known to be sticky across time, and so have persistence built in by behavioural design. The effects of defaults are well documented in the literature (Thaler and Sunstein, 2021; Thaler and Benartzi, 2004; Johnson et al., 2012). From an analysis standpoint, Paper 2 (Chapter 4) engages this point directly by exploiting a multi-year panel to evaluate both the introduction and the withdrawal of a vegetarian-meal subsidy, with event-study analysis tracing the dynamic response. And as I do in that paper, with a large

enough sample, one can partition by sub-group to not only study whether there is persistence in effects but also for whom there may be some long lasting effects.

2.1.3 Mechanisms and channels

A separate but related persistence and credibility concern is whether an intervention is working for the right or expected reasons (Ludwig et al., 2011). Andreoni (1995) show that part of the ATE measured (altruistic behaviour) was not in fact driven by the assumed altruistic preferences, but rather by confusion about the rules of the economic game in the experiment. To understand whether the outcomes observed are actually driven by the intended mechanism, research should strive to measure the channels through which the intervention operates, in addition to its outcomes. And again, this level of mechanism analysis can also help clarify which circumstance or predisposition aspects lead to any observed behavioural heterogeneity. This of course requires micro data, which can be more costly to collect, and is still not common practice today. Haaland et al. (2023); Ludwig et al. (2011) discuss ways of measuring mechanisms, whether information interventions are indeed updating beliefs, or other policy contexts, respectively.

2.1.4 Ethics of Behavioural Policy

The last two decades of growth in behavioural research have been matched by a parallel literature on its ethics. *Even if we can influence people's behaviour for good, should we?*¹ Libertarian paternalism, nudging, and behavioural interventions more broadly share a common feature: the environment in which a decision is made is deliberately designed to shape preferences, perspectives, or choices, sometimes without the individual's awareness, with no transparency (Hausman and Welch, 2010). A common objection rests on the implicit assumption that

¹We do not discuss behavioural interventions employed for private-sector commercial gain, which raise a distinct set of ethical and welfare questions.

the current status quo is fair and neutral, that is, a behavioural intervention introduces bias into an otherwise unbiased environment.

This assumption does not hold. Choice environments are unavoidably never neutral; one option can always be made more salient, more convenient, or selected by default over another, whether by deliberate design or by happenstance (Thaler and Sunstein, 2021; Thaler et al., 2013; Sunstein, 2017, 2015; Thaler, 2018). Nor are individuals the fully sovereign agents of their choices that typical objections assume (Hausman and Welch, 2010). As bounded rational beings with limited attention and self-control (Simon et al., 1972), consumers never have full agency over their preferences and cannot reliably act in their own interests. Were that not the case, the biases on which choice environments operate would not exist, and the debate would be moot.

It is on this point that Sunstein (2017) argues Mill’s anti-paternalism falls short. John Stuart Mill was surely not as acutely aware of the extent to which we are affected by the biases that shape our choices of food, our voting on ballots, and our retirement or health decisions. Mill (1859) states the Harm Principle (or Liberty Principle) as follows:

“That the only purpose for which power can be rightfully exercised over any member of a civilised community, against his will, is to prevent harm to others. His own good, either physical or moral, is not a sufficient warrant.”

A further pushback against the ethicality of behavioural interventions is that they can be untransparently manipulative, with some behavioural interventions exploiting ‘system 1/unconscious’ processes and therefore disregarding aspects of autonomy and the portion of rational agency the consumer does possess (Bovens, 2009; Hausman and Welch, 2010). This lack of transparency, and exploitation of one’s naivety is part of the Nudge+ design, that combines a ‘think’ (system 2) educational element to the nudge (Banerjee and John, 2024).

In the absence of full rationality, moral clarity, and self-control, Mill’s position cannot be sustained (Sunstein, 2017). The ethical question therefore shifts from *whether* to intervene to *how* to intervene well. More specifically, how can policy help people “do the right thing willingly”² while staying on the right side of the ethical line?

Across the discussion of education on our system 1 failings and deliberative thinking suggesting Boosts and Nudge+ interventions (Banerjee and John, 2024; Hertwig, 2017), Lades and Delaney (2022) propose a framework for answering exactly this question of *how* to design Nudges.

The authors propose the mnemonic ‘FORGOOD’ to evaluate the ethicality of a behavioural intervention. The FORGOOD framework consists of seven criteria, Fairness, Openness, Respect, Goals, Opinions, Options, and Delegation.

Fairness considers the distributional consequences of the intervention, and whether it is fair to those affected. For example, Funke et al. (2022) highlight the uneven effective impact of applying a carbon tax on food. They show how this would disproportionately affect the spending budgets of lower-income households, and this is their case for Tax Recycling as an additional policy to dampen the regressive nature of the policy. A closer parallel in the nudge literature is the discussion of opt-out organ donation defaults added to the final edition of *Nudge* (Thaler and Sunstein, 2021). Here the fairness cost is borne not by the target of the policy but by bereaved families, who must actively contest the default when they believe it misrepresents their loved one’s wishes.

Openness ties back to the transparency point made by Bovens (2009); Barton and Grüne-Yanoff (2015) and Hausman and Welch (2010), and queries whether the behavioural policy is open about its intentions, hidden or manipulative. Lades and Delaney (2022); Bovens (2009) suggest two anchor points to decide

²Phrasing borrowed from Weber (2013).

on Openness, i) whether the intervention has been communicated to its recipients, and ii) whether the recipients are conscious of the intervention's influence on their behaviour. Other proponents of this attribute defend openness, such as through public announcements of policies, as a self-regulatory mechanism, ensuring that only policies that policy-makers are willing to defend publicly under public scrutiny are applied (Thaler and Sunstein, 2021).³

The Respect attribute questions whether the behavioural intervention respects the autonomy of the individuals involved. For example, behavioural interventions informed by Type 1 decision processes affect behaviour through automatic, unconscious processes (Kahneman, 2011), and from this perspective could be considered as more restrictive to the agency of the individual.

Goals and Opinions revolve around taking an overview perspective on the objective of the behavioural intervention, and evaluate whether the policy is doing genuine good and whether the public support this policy with the goal in mind. Options and Delegation are a further due diligence of the policy, and assess whether better alternatives to achieve these goals exist, and importantly, whether the policy-maker is empowered to incite such behavioural interventions and intervene on behaviours of the target population in the first place.

The Options and Goals criteria connect to a broader ethical critique of the behavioural policy enterprise itself. Chater and Loewenstein (2023) argue that behavioural interventions, by virtue of being individually targeted ('i-frame') and comparatively cheap to adopt, act as low-hanging fruit that can distract from larger regulatory or structural ('s-frame') reforms that might be more effective but are costlier and more politically difficult to implement. On this view, the ease of adoption of behavioural policies is not only a practical advantage but an ethical liability for wider society, insofar as it shifts attention and potentially capital

³This is also discussed as the 'Publicity Principle' (Lades and Delaney, 2022).

away from more consequential structural alternatives. Most proponents respond that behavioural interventions were never intended to replace structural policy, but to complement it. Behavioural insights can sharpen the design of structural policies (e.g. default architectures, tax salience (Chetty et al., 2009)) and help anticipate the likely behavioural response to them (Marteau et al., 2025; Connolly et al., 2025; Benartzi et al., 2017). This can be useful to consider when evaluating Options and Goals under FORGOOD, or general behavioural policy suggestions, since it asks the researcher to consider not only whether the intervention works, but whether pursuing it forecloses a more ambitious policy path.

2.1.5 Generalisability and Scalability

Overview Scalability and external validity are another important set of hurdles in behavioural policy research (Al-Ubaydli et al., 2017). Even if a finding is replicable, it may not survive when the intervention is deployed at policy-relevant scope. An intervention that works in a small-scale study may not have the same effect when implemented across a larger population, due to a variety of factors including the upward-biased results discussed in Section 2.2.2, a more heterogeneous population when scaling up, or spillover effects from other parts of participants' lives that are not absorbed in a study's setting. DellaVigna and Linos (2022) illustrate this with the discrepancy in effect sizes ('M-effect') between nudge studies and their Nudge Unit Randomised Control Trial (RCT) equivalents. The gap is in part driven by the same factors that drive the replication gap, but also by the fact that the Nudge Unit RCTs are run in real-world settings with a larger and likely more heterogeneous population (Al-Ubaydli et al., 2017). Another reason is that most experiments are partial equilibria, e.g. only study the effects of an intervention from a consumer perspective, not the response from the supplier who will undoubtedly have an effect on the overarching, total equilibrium, outcome.

Research results being considered for real-world application, in our case for be-

havioural policy, therefore need to be interrogated on how they will scale, whether they have external validity, and whether their estimated and take-away results are generalisable (Al-Ubaydli et al., 2017). While piloting policy interventions as RCTs in the real-world would be the ideal, this would be costly and administratively infeasible if several papers compete for the same policy solution spot. Although RCTs also have their critiques, mainly following the lines that scaling up “gold-plated” RCTs will not have the same effects as the original RCT as the latter will be the first to introduce the less than perfect participants for the study (Duflo, 2020; Deaton and Cartwright, 2018; Pritchett and Sandefur, 2014).

As an aside, large research institutes circumvent some of the costliness, not the sampling bias, by running several RCTs in tandem (known as Megastudies) (Milkman et al., 2021) and selectively dropping treatments during interim analysis, but this is incredibly expensive and not common in academic research. This methodology also helps with between treatment comparisons, as these will be with similar experimental designs, similar population, same background period, etc.

Other frameworks have therefore developed sets of methods to evaluate the scalability and generalisability of behavioural interventions. For example, List (2020) proposes a formal framework of conditions under which experimental results are more likely to be generalisable to a real-world setting. The proposed ‘SANS’ framework is composed of four criteria, and by verifying that these criteria are satisfied, it is argued that an experiment’s ATE is more likely to be generalisable to the policy-relevant ATE at scale.

SANS framework. The first criterion is *Selection*. It queries whether the selection of participants in the experiment is representative of the policy’s target population. Sample bias can come from running experiments with students, WEIRD (Western, Educated, Industrialised, Rich, Democratic) populations, or volunteer participants; or from field experiments where the host can self-select

into the experiment (so that the ATE is based on a more sensitive portion of the population). This is in line with site selection bias arguments of Allcott (2015). The caveat in each case is that these are not the target populations of the policy.

In a similar vein, *Attrition* queries whether there is differential, heterogeneous drop-out of participants across treatment and control groups, which will bias the results one way or another. For example, only those willing to be affected by an intervention continuing the study over time, leading to an upward biased ATE.

Naturalness considers how similar the experimental context is to the policy setting, not only in the background and functionality, but also in the incentives and other economic mechanisms (e.g. effects on others, spillovers to other aspects of life). There is also a query with regards to the Experimenter Demand Effect (Zizzo, 2010) (Hawthorne Effect). This is part of the argument in favour of field experiments, which are more naturalistic and where participants are not aware of being observed. Framed field experiments are an option for the natural setting, but often the participants are made aware of and have consented to participation in the study (although the study might not necessarily be explicitly detailed ex-ante, avoiding some self-selection and experimenter demand effects). These pros and cons are also discussed in Harrison and List (2004).

Finally, *Scaling* queries how general equilibria might change when an intervention is applied at scale. For example, changing behaviour at scale might change social norms, and hence have a different effect than what is observed in a small-scale experiment. Theory by Mattauch et al. (2022) details this catalytic effect of a policy affecting large-scale behaviour, changing the social norm, such that endogenous norm-based preferences subsequently feed back into behaviour, creating a second layer of effects, a positive feedback loop of sorts.⁴ In Paper 3 (Chapter 5), my

⁴In that paper, the authors argue that, to study the effects of climate policy, understanding whether endogenous preferences will crowd out or crowd in the policy is crucial to measuring its full effects and value.

paper includes a SANS analysis, and I further discuss the SANS criteria in the Discussion Chapter (Chapter 6) in relation to my own work.

Issues with adoption. Beyond these, there has been a more recent push to incorporate RCTs into the process of policy innovation (Baron, 2018), verifying potential results in the field before scaling. These results, their strength, low cost, and low barriers to adoption, are not necessarily the predictors of policy adoption, against expert forecasting (DellaVigna et al., 2024). DellaVigna et al. (2024) suggest that organisational inertia is often a large bottleneck and biasing factor for which policies are adopted. The authors consider RCTs done in collaboration with Nudge Units that involved communication interventions, not structural policy changes, so they are all low cost. Interventions that only required a change to an already existing communication (an additional sentence, a change of framing, etc.) were more likely to be adopted than an intervention that required a whole new communication to be set up.

As discussed above with DellaVigna et al. (2024), the conversation around moving from research evidence to policy may not only be about the strength and cost of the intervention, but also about how smoothly the change can be implemented within existing organisational and political structures. This is an extension in reasoning from just what works to also what gets adopted. Some might argue that this should not be viewed as an extension or second step, but as a precursor to testing the strength of the intervention. However, I think that might be selling society short. Science must highlight and study the best-case scenarios in addition to the more realistic ones, and if we only study the latter, we might miss out on some of the most powerful interventions that could be adopted with a little bit of organisational change.

Moral licensing, crowding-out and spillovers. A related scalability concern, particularly in pro-social behavioural policy, is that the headline effect may be undone by offsetting responses, either within the same domain through moral licensing or motivational crowding-out (Gneezy and Rustichini, 2000; Mullen and Monin, 2016), or across domains through spillover into other parts of the individual’s life. The common driver is that the underlying moral identity of the person has not been changed, and so an overly good action provides excess self-signalling credits which a person can ‘spend’ by being a little bit more badly behaved while maintaining their identity level (Bénabou and Tirole, 2011, 2006). This moral licensing could also occur across domains, through spillover effects (Gneezy and Rustichini, 2000; Mullen and Monin, 2016). These offsetting responses are particularly difficult to detect in field experiments, where the researcher will generally not have access to wider life behaviour data of the individuals beyond the direct study variable. I argue this reinforces the importance of measuring the channels and mechanisms affected as well as the intervention with a longer time frame. If, for instance, a nudge does not move (self-image) beliefs, then moral licensing is more likely to occur than if the nudge does encourage belief updating.

Cost-effectiveness. Cost-effectiveness of the policy intervention is particularly important for behavioural policy, where the effects are often smaller than those of traditional policy instruments, but also much cheaper to implement. For example, a nudge can be effectively costless per person, whereas a subsidy can be financially and administratively costly to implement. Cost-effectiveness will also depend on whether effects persist and are stable during and beyond the policy (Section 2.1.2), and on whether the intervention has positive or negative spillovers to other domains. Benartzi et al. (2017) argue that the cost-effectiveness of nudges should be reported in terms of the effect per dollar spent, and compared with traditional policy instruments on the same basis, although we need to

remember that historical effect sizes have been over-amplified (DellaVigna and Linos, 2022). Beyond the cost-effectiveness of the intervention itself, the cost of organisational change to adopt it should also be considered, in relation to already existing infrastructure (DellaVigna et al., 2024).

2.2 Methods for Behavioural Policy

Having considered the distributional, dynamic, ethical and generalisability consequences of behavioural heterogeneity, I now turn to how we go about identifying and measuring the effects in the first place. This section gives an overview of the behavioural policy researcher’s methodological toolkit.

2.2.1 Study Environments

To start off a behavioural policy research project, even perhaps before committing to a specific research question and intervention, the researcher should consider the design of and the identification strategy possible. This includes data collection, time frame, ethical feasibility and budget. These are the practical constraints often more prominent in the early stages of research, and they will shape the research question, its identification cleanness.

Laboratory experiments and internal validity From these considerations the behavioural researcher will choose from a range of study environments (Harrison and List, 2004; Levitt and List, 2009; Duflo, 2020). They could opt for a lab-based experiment (online or in-person), these often benefit from being ‘clean’ study environments with no environmental and non-design features affecting behaviour. These typically have the best opportunity for internal validity. However, these also suffer from a lack of naturalness and generalisability, and are more vulnerable to experimenter demand effects and biased participant pool as they

have volunteered for the specific experiment. These are an important stage for researchers trying to test the effects of mechanisms.

Halfway house to field experiments The next closest option from the lab experiment is the lab-in-the-field experiment, which is particularly popular in Development Economics, where the clean laboratory task experiment is run with the desired policy-targeted population and perhaps not in a lab experiment building but a more natural setting but still with a more controlled choice environment, this circumvents the issue of most research being based on WEIRD populations (List, 2020), but it still suffers from the same issues of experimenter demand effects and lack of naturalness.

Another option, which seeks more external validity, is the framed field experiment, where the experiment is run in a natural setting with a more natural population, but the participants are aware of being part of an experiment in which they have consented to participation. This is a common design for behavioural policy research, as it allows for more naturalistic behaviour while still maintaining some level of control and internal validity, compared to a full field experiment. It is also often logistically cheaper, without as much bureaucracy to set up. Whether it is financially cheaper depends on the cost of creating a fake near real framed environment.⁵

Full field experiments Finally, the field experiment is the most naturalistic and has the highest external validity, but it is also likely the most logistically difficult to set up, and it is more difficult to maintain internal validity. It is also more difficult to measure channels and mechanisms in a field experiment, as the researcher has less control over the environment and the participants. And often this type of research looks at aggregate results rather than individual-level

⁵Paper 3 in Chapter 5 is an example of a framed field experiment using an online experiment supermarket environment, Woods, created to mimic a British supermarket environment, like Tesco's or Sainsbury's.

data, which makes it more difficult to understand heterogeneity in effects, and the mechanisms through which the intervention operates.

It is also worth noting, in the behavioural research space we need not have deliberate field experiments, as RCTs, but often we can also work with natural field experiments, where the researcher did not design the policy intervention, but can take advantage of a policy change to extract causal effects from that change in circumstances.

More discussion on these lab and field settings include Harrison and List (2004), Duflo (2020) and Levitt and List (2009).

2.2.2 Replicability

It is useful to first distinguish two related notions. Reproducibility typically refers to obtaining the same results from the same data and analysis, while replicability refers to obtaining consistent findings in an independent study with newly collected data that addresses the same question. As experimental and secondary-data research has become more popular, both of these concerns have come to the foreground (Ioannidis, 2005; Collaboration, 2015; Camerer et al., 2019). The concern is especially sharp for behavioural policy research, where single-study findings can be promoted into real-world policy design with a faster turnaround and fewer subsequent opportunities for the broader literature to stress-test the underlying claim, compared with the typical science of human behaviour, where several investigative papers usually debate the claim before any policy traction is gained.

The interest in replicability was pushed to the forefront of debate by the Reproducibility Project: Psychology (RPP) (Collaboration, 2015) and the Experimental Economics Replication Project (EERP) (Camerer et al., 2016), both of which used large samples to mount more highly powered replication attempts of their

selected publication results. The replication success rates were much lower than expected, with only 36% of the RPP's 100 replications and 61% (11 of 18) of the EERP's replications successfully replicating the original findings, indicative of a large proportion of false positives in the original papers. Both projects also examined effect sizes. Among studies that replicated with the same sign, magnitudes were typically much smaller than the originals, with the RPP finding a median replication effect size around half of the original. Camerer et al. (2016) extend this exercise to social science papers published in *Nature* and *Science*, general-interest outlets that tend to publish high-impact, headline findings, and document the same pattern of directional and magnitude bias. These results are consistent with the better-known statistical phenomena of Type-M (magnitude) and Type-S (sign) errors that arise when underpowered studies are filtered through statistical significance (Gelman and Carlin, 2014).

Tests of the replication gap. The replication-rate figures should nevertheless be read with some care. The replication projects themselves have attracted methodological pushback, for example over the fidelity and representativeness of the chosen replications. Camerer et al. (2019) note in particular that a potential confound is the use of Amazon Mechanical Turk (AMT) participants, whose repeated exposure to similar social science tasks could affect results.⁶ This concern is part of a broader point about learning and confusion in experimental tasks. Andreoni (1995), for instance, argues that only about half of the apparent altruistic behaviour observed in public good games is genuinely altruistic, with a large share of cooperation driven by errors and misunderstanding of the game. This underscores the importance of learning and comprehension checks in experimental design, and, more broadly, of measuring the channels through which an intervention operates rather than its outcomes alone. I discuss this point fur-

⁶Camerer et al. (2019) raise this in response to a specific criticism of their replication methodology rather than as a central point. They also find that two of the three economic games they consider replicated successfully.

ther in Section 2.1.2 in relation to my own work and the importance of studying channels and mechanisms of effects.

The replication gap is also visible in behavioural policy itself. DellaVigna and Linos (2022) compare published academic estimates of nudge effects with the larger-sample Randomised Control Trials (RCTs) run by Nudge Units, finding that, while the direction of effects is broadly preserved, the effect sizes in the higher-powered field RCTs are smaller than those in the academic literature (around one-sixth on their headline estimate). They attribute part of this gap to publication bias, the tendency of journals to publish positive and statistically significant findings rather than null or negative results (Ioannidis, 2005; Collaboration, 2015; Rosenthal, 1979), an issue we do not treat in depth here.

Beyond publication bias, three further drivers of inflated false-positive rates are routinely identified (Ioannidis, 2005; Collaboration, 2015). The first is low statistical power, which leads to a higher rate of false positives and inflates the magnitude of those that are detected. The second is p-hacking, where researchers try different regression specifications, control sets, and outcome definitions and report those that cross conventional significance thresholds. The third is broader analytical flexibility. For example, Silberzahn et al. (2018) show that 29 independent teams analysing the same dataset and addressing the same research question reach materially different conclusions, illustrating that even with identical data the choice of specification can drive the result.

Pre-commitment documents. Several remedies have been proposed and partially implemented in branches of academia. The first group targets ‘researcher degrees of freedom’. Pre-analysis plans and pre-registration, in which researchers specify their research question, hypotheses, data collection, and analysis plan before collecting data, tackle p-hacking and analytical flexibility (Ioannidis, 2005; Collaboration, 2015; Camerer et al., 2019), and several journals now require pre-

registration on a recognised repository as a condition of editorial consideration and acceptance. The mechanism is however imperfect. There is as yet no strong consensus on the level of detail required, with an incentive to remain as general as possible, and no mechanism prevents pre-piloting of experimental ideas before registration. Registered reports go further, peer-reviewing and accepting the paper on the basis of question and methodology before data collection, and primarily target publication bias (Chambers and Tzavella, 2022). A second group is for prediction markets and expert survey forecasts which have been shown to be reasonably accurate predictors of replication success (Camerer et al., 2019; Dreber et al., 2015; DellaVigna and Pope, 2018).

These remedies raise the credibility of the published literature, but it is worth asking whether their complete adoption risks cutting the space for exploratory, high-innovation and risk research. If pre-analysis plans commit us to a narrow set of hypotheses and secondary analyses, the field may lose some of the (speed to) surprising findings. If prediction markets are accurate, that seems a poor indictment of research progress and innovation. There is surely a trade-off between stable, well-disciplined science and the more exploratory work that produces large conceptual jumps, and it seems clear that a healthy academia needs both.

2.2.3 Identification Strategies

The set of econometric and statistical tools for identifying causal effects of behavioural interventions is large and growing, and are assumed to be known, so will not be discussed in detail here. Instead, I discuss some of the more recent contributions and debates in the literature, of relevance to behavioural policy research.

Difference-in-difference regression Although cross-sectional regressions remain commonplace, difference-in-differences (DiD) designs have become the

workhorse for causal analysis in behavioural policy research (Roth et al., 2023; Bertrand et al., 2004). Part of their appeal is that they are practical for when a randomised intervention is not feasible, DiD allows the researcher to exploit existing policy changes as natural field experiments. Even for experiments, these have methodological advantages. DiD relaxes the strong comparability requirement of cross-sectional work. Rather than assuming that treated and untreated groups are composed of perfectly comparable people, DiD only requires that, pre-intervention, their outcomes would have evolved in parallel, which can be somewhat tested (Roth et al., 2023). By differencing out time-invariant aspects (e.g. socio-demographics or individual tastes and preferences) and common time shocks to all groups, the estimator therefore absorbs some of the issues with cross-sectional comparisons of non-equivalent populations.⁷

Difference-in-difference has two other interesting features when compared with cross-sectional analysis, and these further features make DiD particularly well suited to behavioural policy. First, because identification comes from within-unit variation around a policy change (following a person's before and after intervention behaviour change, ideally considering any seasonal and other cyclical effects), longer panels make it possible to study the persistence or decay of effects beyond the short horizons, a feature that Paper 4 exploits directly. This is important as nudge interventions effects have been shown to fade over time in some cases (Allcott and Rogers, 2014). One explanation can be the salience of the change which then fades as it is normalised in our memories (Bordalo et al., 2020). Second, difference-in-difference with repeated data points per person can offer much more statistical power and precision for a smaller sample size, and so offers opportunity to explore heterogeneity in effects across different types of individuals.⁸

It is of course worth the caveat that difference-in-difference studies are more

⁷Which then require several additional balance, heterogeneity and robustness checks.

⁸It is worth reminding the reader that often in research projects, the research budget is a primary constraining factor and so sample size is often imposed, not chosen.

susceptible to experimenter demand effect, or some over-reaction to studies as treated participants receive both the control (pre-treatment) and the treatment (post-treatment) manipulation. However, this is also reflective of the reality of a real-world policy change, where the population will be exposed to both the pre- and post-treatment environment, so it is not necessarily a problem for external validity. From an empirical analysis standpoint, the longitudinal data yields potential bias in error terms and for this discussion I leave the reader to start with seminal works such as Bertrand et al. (2004); Colin Cameron and Miller (2015); Abadie et al. (2023).

Caveats with DiD A more recent concern with DiD that is particularly pertinent to behavioural policy is that, under staggered adoption and heterogeneous treatment effects, the conventional two-way fixed-effects (TWFE) estimator can produce a biased estimate of the average treatment effect, and in extreme cases an estimate of the wrong sign (Goodman-Bacon, 2021; De Chaisemartin and d’Haultfoeuille, 2020; Callaway and Sant’Anna, 2021).⁹ The intuition is that TWFE recovers a weighted average of all available 2×2 comparisons in the panel, and some of those comparisons use already-treated units as the control group for later-treated ones. For clarity, if a policy is rolled out to one region in January and another in June, the January cohort is treated as a control when identifying the June effect simply because it is not switching status at that moment. This is not a major concern only if the January cohort’s outcomes are swiftly stable post-treatment, and settled by June, so that the only changes in behaviour being captured for that cohort from June onwards are the same as the ‘true’ control’s (e.g. seasonal effects). On the other hand, if the effect of the treatment on the early group is still growing or decaying, those changes will be considered as control changes and incorrectly subtracted from the June cohort’s

⁹Fixed effects are the standard device for sweeping out time- or unit-invariant factors from the treatment estimator, for example individual tastes, seasonal effects, or country-level norms and legislation.

change.

These conditions are exactly those one expects in behavioural policy. Rollouts are routinely staggered across regions and time.¹⁰ With regards to the stability of treatment effects, evidence and theory suggest that nudge effects in particular often fade once the intervention loses salience (Allcott and Rogers, 2014; Bordalo et al., 2020). Several staggered adoption DiD techniques have been recently proposed, using only never- or not-yet-treated units as controls, and then later aggregating these estimated effects among others (see Callaway and Sant’Anna (2021); De Chaisemartin and d’Haultfoeuille (2020)).

Other Policy Regression Methods Regression discontinuity (RD) is another quasi-experimental design that can also be interesting for causal identification in behavioural policy research. Its design and use revolves around policy treatment being assigned based on a cutoff rule, for example earning below a certain income threshold, or living within a specific radius, etc. RD exploits the sharp step in treatment assignment, to compare those just above and just under the cutoff, assuming that these individuals are in every other way comparable. For example, if the cutoff is for those born on 14th of November, those born on a few days prior are probably not different from those born a few days after. Similarly if the policy is for those within 40 miles of a city, those living in the doughnut area with a couple of miles each side might not be so different. This type of regression requires a large population closely placed around the cutoff, and these strict cutoffs are not that common, or easily known about.

Event studies are another recent popular analysis tool, generally as part of wider analysis and to specifically test dynamical effects of interventions (how long an intervention takes to take effect, whether it fades over time, etc.) (Schulze-Tilling,

¹⁰As opposed to the cleaner case of a single treatment group receiving the intervention at one point in time, with a straightforward before-and-after comparison against an untreated group.

2025; Epperson and Gerster, 2026).

A final note on new identification techniques is presented for climate policy effects (Stechemesser et al., 2024). These authors reverse engineer the identification problem by first studying a large dataset for effects ('break-detection') and then evaluate whether that break can be attributed to a policy change that occurred within a credible timeframe from adoption to effect. This is a useful approach for climate policy, where the outcomes (emissions output) will be delayed. This however requires large datasets, large effects or breaks to detect in the first place, and a large number of same policies to have been adopted in different locations and times, so that the break can be credibly attributed to a policy change rather than some other region specific shock.

Chapter 3 | Paper 1

Should I Stay or Should I Go: Intentions to Freelance Across 16 Countries

Abstract: A substantial literature explores entrepreneurial self-employment, whereas freelancing, a distinct and increasingly prevalent form of self-employment, is less well studied. Using an original large survey conducted in 16 countries, we investigate the key determinants of intentions to engage in these activities. We confirm the importance of socio-demographic characteristics and specific behavioural factors, but, as expected, some findings differ from the entrepreneurship transition literature (Simoes et al., 2016; Evans and Leighton, 1989). A detailed heterogeneity analysis highlights that intentions to engage in freelancing varies across job and individual contingencies. Of concern, vulnerable cohorts have higher interest in this form of work with lower worker protections and higher risks.

Declaration of Authorship: I declare, as per University of Oxford School of Geography and the Environment regulations, that I have contributed more than the required 50% to the intellectual content of this paper, and that I am the primary author.

Authors: Adam Ferris and Stefania Innocenti.

The paper is included in the next pages.

Should I Stay or Should I Go: Intentions to Freelance Across 16 Countries

Adam Ferris, Stefania Innocenti

Abstract

A substantial literature explores workers transitions into entrepreneurial business ownership, while freelance, on-demand self-employment, a distinct and increasingly prevalent mode of labour market participation, remains comparatively understudied. Using an original large survey conducted in 16 countries, we investigate the determinants of workers' intentions to leave their main salaried job to engage in freelance, on-demand self-employment. We find that socio-demographic characteristics, job conditions, and behavioural traits are systematically associated with these intentions, though several patterns diverge from those documented in the entrepreneurship literature. A detailed heterogeneity analysis substantiates variation across occupations, task types, and individual characteristics. Workers in more vulnerable positions exhibit higher intentions to transition into freelance, on-demand self-employment, a form of work often associated with lower protection and higher income risk.

Non-Technical Abstract: Who is expressing interest in freelancing? Using survey data from workers in 16 countries, this study examines which socio-economic, demographic, psychological, and job-related factors are associated with an interest in leaving traditional employment to engage in freelance, on-demand work. By comparing patterns across countries and worker groups, we show that freelancing differs in important ways from other forms of self-employment that are often studied together in the academic literature. These results underline the importance of distinguishing between different forms of self-employment and point to the need for targeted policy responses as freelance and on-demand work continues to expand.

Keywords: Freelancing, Self-employment, Labour Policy, On-demand

JEL Codes: L24, L26, J24, D10, M21

1 Introduction

A considerable body of research has examined transitions from wage employment into self-employment, with a particular focus on entrepreneurial business ownership. (Rees and Shah, 1986; Kolvereid and Isaksen, 2006; Woronkiewicz and Noonan, 2019; Guerra and Patuelli, 2016; Ekelund et al., 2005; Stewart Jr et al., 1999; Hansemark, 2003; Simoes et al., 2016; Wyrwich, 2013; Evans and Leighton, 1989; Blanchflower and Oswald, 1998; Kim et al., 2006). In much of this literature, entrepreneurship is conceived as a specific form of economic activity centred on opportunity recognition, innovation, and business creation.

However, over recent decades, new and more fragmented forms of self-employment emerged. The expansion of the internet has acted as a significant catalyst, enabling more dynamic and diverse portfolios of work¹ including project-

¹In his pioneering work in organisational behaviour, Handy (1984) described a shift towards portfolio careers, whereby firms outsource tasks and individuals combine multiple work engagements.

based and on-demand forms of self-employment, rather than traditional linear career progression, or business ownership. Recent evidence suggests these arrangements are growing in the world of work, as full-time, part-time or ‘side hustles’ (van Gelderen et al., 2021; Kitching and Smallbone, 2012).

Importantly, these forms of work are not homogeneous. On the one hand, part of the literature emphasises precariousness, weakened labour protections, and power asymmetries, particularly for lower-skilled workers engaged in task-based on-demand activities (Prassl, 2018). On the other hand, project-based self-employment may also take the form of professional contracting, through which firms hire specialised expertise on a temporary basis, often involving relatively high-skilled workers with higher remuneration. As noted by Henrekson and Stenkula (2010), self-employment is neither a necessary nor a sufficient condition for entrepreneurship, and the intended function of self-employment is a key distinguishing factor across different role types.

In this paper, we study intentions to transition from salaried employment into own-account, freelance self-employment characterised by on-demand, project-based work, rather than entrepreneurial business ownership. This definition includes certain forms of on-demand work - such as task-based or platform-mediated engagements - but is not limited to them. Overall, however these activities are typically governed by fixed, short-term contracts (Prassl, 2018; van Gelderen et al., 2021).² In these contexts, workers typically face low entry barriers, make limited upfront investments, and do not pursue the development of a personal business concept into a long-term entrepreneurial opportunity. As a result, engagement in these activities may be viewed as a flexible means to secure income, autonomy, or short-term career adjustment rather than as an investment-driven

²On-demand work can be categorised as *crowdwork* if tasks can be completed solely online (e.g., copy-editing or accounting), or as *gig work* if tasks must be performed offline. This distinction remains even if the task originated from an online platform. Examples include Uber and Deliveroo drivers hired via digital platforms, as well as independent contractors such as plumbers and plasterers engaged by project managers.

career path (Scarelli and Margolis, 2023).

While freelance, on-demand self-employment offers several potential benefits, including increased flexibility, opportunities to pivot careers rapidly, and the ability to supplement primary income streams, it might also come at the expense of social protection rights. Depending on the institutional and legal context, workers may fall outside the scope of traditional employment protections or be expected to self-insure against risks typically covered in salaried work (Prassl and Risak, 2015; Prassl, 2018). The extent and nature of these risks vary substantially across countries, reflecting differences in labour law, social insurance systems, and regulatory approaches. These cross-country differences motivate a comparative perspective.

Although the costs and benefits of business ownership have long been debated, the rise of the on-demand economy has heightened the relevance of these discussions. Existing research has extensively explored how individual characteristics, such as demographic background, family circumstances, job conditions, and personality traits—influence transitions into conventional self-employment, most often entrepreneurship and typically within a single country or industry.³ However, while decisions to enter entrepreneurship and own-account, non-business-owning self-employment likely share some common determinants, they are also distinct in important aspects (Burke and Cowling, 2020a; Kitching and Smallbone, 2012). Yet empirical evidence on the determinants of workers’ decisions to enter project-based own account work remains limited, largely due to data constraints and inconsistent definitions (Audretsch et al., 2015; Kässi and Lehdonvirta, 2018).

This paper contributes to filling this gap by investigating workers’ intentions to engage in freelance, on-demand self-employment. Our dependent variable captures the workers’ stated intentions to leave salaried employment to pursue free-

³See Simoes et al. (2016) for a review. See also Hansemark (2003) for an investigation of locus of control, Guerra and Patuelli (2016); Stephan et al. (2020); Nikolaev et al. (2020) for job welfare, Wyrwich (2013) for cultural background and Kihlstrom and Laffont (1979); Rees and Shah (1986); Stewart Jr et al. (1999); Ekelund et al. (2005); Caliendo et al. (2009) for theoretical and empirical research on the role of risk tolerance.

lance work within the next 12 months. As such, our analysis focuses on intended transitions into freelancing rather than realised labour market transitions. Using harmonised survey data from approximately 18,000 workers across 16 countries, we investigate the socio-economic and psychological characteristics associated with individuals' propensity to consider transitions into freelance work.

We first document heterogeneities in workers' intentions to enter freelance, on-demand self-employment. While 17% of our survey participants overall are willing to voluntarily leave their current occupations to engage in freelance work, there is a 30-percentage point difference across the 16 countries in our sample.⁴ Intentions also vary significantly across job task-types (Acemoglu and Autor, 2011) and occupations.

Second, we estimate a model of intentions to take up freelance work. With no claim of causality, the results suggest that older, female, more educated and higher-income people are less likely to consider moving to freelance, on-demand self-employment in the next 12 months, diverging from patterns commonly found in the entrepreneurship literature. We also find evidence of non-linear relationships between age, education and intentions to freelance. Considering current job characteristics, the number of past jobs held, working in creative roles, atypical work arrangements⁵ or private sector employment are positively associated with intentions to enter on-demand, freelance work arrangements.

Finally, we find that behavioural traits play an important role. Higher risk aversion is associated with lower intentions to freelance. Lower scores in the cognitive reflection test (CRT), correlated with higher impulsivity and impatience, are associated with higher interest in freelancing, while positive reciprocity also increases the likelihood of considering own account, project-based work.

⁴See Figure 2.

⁵The definition of non-standard forms of employment encompasses all forms of employment other than a full-time permanent contract (Eurofound, 2017). The same definition is used in recent work on atypical work arrangements - see for example Datta et al. (2019) or Innocenti and Golin (2022).

Our contribution to the literature is two-fold. First, we build on previous literature on behavioural labour economics which has typically been interested in topics related to entrepreneurship, but has, so far, not investigated workers' propensities to engage in freelance, on-demand self-employment work with a clear homogeneous definition (Audretsch et al., 2015).

Equally, whilst most studies concerned with self-employment decisions, with a focus on business ownership, have focused on a limited number of countries or age cohorts (Blanchflower and Meyer, 1994; Lin et al., 2000; Borjas, 1986; Stephan et al., 2020), our original large survey conducted in 16 countries with representative samples of workers aged between 20 and 70, allows for empirical analysis on a wide set of geographies and population segments. This allows us to better account for cultural, economic and jurisdictional differences affecting intentions to entry own-account, project-based work.

Overall, our findings underscore the importance of treating freelancing as a distinct category within self-employment research (Audretsch et al., 2015) and highlight the need for policy attention to workers transitioning into freelance work and the potential long-term societal implications of this growing labour market segment (Burke and Cowling, 2020a,b).

2 Survey Design

We rely upon a large international survey whose responses were collected between February and March 2019. The research work presented in this paper employs the same survey described in Innocenti and Golin (2022); Innocenti et al. (2023). The data originate from an international study launched by Zurich Insurance through a survey administered online by the Dutch professional survey company, Epiphany. Individuals aged between 20 and 70 years old and currently residing in Australia, Brazil, Finland, Germany, Hong Kong, Ireland, Italy, Japan, Malaysia,

Mexico, Romania, Spain, Switzerland, the United Arab Emirates (UAE), the United Kingdom (UK) and the United States (US) were selected to participate in the survey. For each country, hard quotas for age and gender were set to ensure representativeness along these characteristics.⁶

The survey company recruited respondents through several panel providers, enabling global outreach, as well as access to local recruitment expertise. Recruitment followed jurisdiction-specific data protection and privacy laws. Compensation arrangements, for those who completed the survey, were set based on each panel provider’s norm. The possible compensations included points-programmes, gift cards, vouchers, charitable contributions and prize draws.

The survey was primarily designed to capture information on whether and how workers were adapting to different facets of the labour market, and its fast-paced changes. To enable multiple lines of enquiry, the survey was comprised of several sections which focused on different aspects of the respondents’ relationship and interactions with the world of work, amongst other individual nuances, characteristics and traits. We describe the relevant overarching sections below.⁷

Intentions to freelance Respondents were also asked about their willingness to take up freelance work which represents our outcome variable. More precisely, respondents were asked to report their likelihood of leaving their current main job within the next 12 months in order to become a “freelancer (become self-employed and work on demand)”. This question followed an earlier survey item that classified respondents’ current employment status and described self-employment as

⁶In particular, age and gender quotas were implemented as hard quotas agreed ex ante and enforced by the survey provider within each country. For 13 countries, the age quotas were based on OECD data which captured employment status. For three countries, the age quotas were set based on World Bank (National census) data but did not consider employment status. In addition, the survey aimed to ensure a broad regional spread using soft quotas, to avoid excessive geographic concentration of respondents. For two countries, additional context-specific soft quotas were applied: respondents’ country of birth in the UAE and respondents’ ethnicity in Malaysia.

⁷The exact formulation of all questions used in this paper is provided in order of presentation in Appendix C.

an umbrella category encompassing either owning one’s own business or freelancing/working on demand. As such, the outcome variable used in this paper explicitly focuses on intentions to enter a form of self-employment characterised by project-based freelancing combined with on-demand engagement, thereby excluding transitions into traditional business ownership or entrepreneurial self-employment.

Answers were given on a four-points Likert scale ranging from 1 “Very unlikely” to 4 “Very likely”.⁸ in this form of freelance, on-demand self-employment, and zero otherwise.⁹ Henceforth, for brevity, we refer to this outcome as “intentions to freelance” or “consider freelancing”, “intentions to work on-demand” with the understanding that this term denotes freelance, on-demand self-employment as defined above.

Additionally, we note that we measure workers’ intentions rather than actual behaviour. The cross-sectional nature of the survey prevents us from being able to internally validate the intentions to freelance question against actual (future) uptake of tasks. However, the use of entrepreneurial intentions are commonly used as proxies of action and seen as relevant information in the field (Bird, 1988; Krueger Jr et al., 2000).

Work situation The survey collected information on respondents’ current work situation, including employment status (with the self-employed and unemployed excluded from the analysis), job mobility (number of past jobs), multiple job holding, private sector employment, and atypical work arrangements. Other questions focused on whether the individual currently worked in the private sector and was subject to an atypical work arrangement. Furthermore, the survey also contained information on the nature of the tasks that workers are conducting. We followed the Acemoglu and Autor (2011) method and constructed a measure of task con-

⁸See Survey Questions in Appendix C

⁹The explicit 12-month time horizon was included to strengthen the link between stated intentions and subsequent behaviour.

tent from their self-reported answers about the prevalence of manual vs knowledge and creative vs routine work in their current job. Specifically, we utilise responses to two questions that asked participants to indicate what percentage of their work they consider to be “manual or knowledge based” and “creative or routine”. If the response to the first (second) question is more than 50%, we classify the work as manual (creative); if not, we classify it as knowledge-based (routine). The four task types that arise from this are knowledge-routine, knowledge-creative, manual-routine, and manual-creative.

We hypothesise workers’ current job characteristics to be systematically related to intentions to freelance. Greater job mobility, multiple job holding, atypical employment, private sector employment, and creative or non-routine task content may lower switching costs into freelance work and are therefore expected to be positively associated with intentions to engage in freelance, on-demand self-employment.

Attitudes and behavioural traits The survey elicited standard behavioural measures of risk tolerance, locus of control, positive reciprocity, and cognitive reflection. Risk attitudes were measured using an adapted investment task following Charness and Villeval (2009) and Charness and Gneezy (2010), in which respondents chose how much of a hypothetical \$100 to invest in a risky asset that yields 2.5 times the amount invested with 50% probability, and zero otherwise.¹⁰ We construct a measure of risk aversion based on the amount not invested and classify individuals as more or less risk averse relative to their country median.

Locus of control was measured through two standard questions adapted from the World Value Surveys and asked individuals on a 5-point Likert scale to indicate the extent to which they believe they have control on their own lives. These questions have in the past been used as proxy locus of control by Veenhoven

¹⁰While behaviour in this task may capture broader behavioural attitudes, including aversion to losses, we interpret this as a risk attitudes measure.

(2000); Inglehart et al. (2008); Verme (2009); Williamson and Mathers (2011); Innocenti and Golin (2022). Answers to both questions range from 1 “No control at all” to 5 “A great deal” of control. We take the average of the answers to the two questions, as done by Veenhoven (2000); Innocenti and Golin (2022), and classify individuals as having an internal (external) locus of control if their value is above (below) their country’s median.

Positive reciprocity was measured using a hypothetical donation task (Falk et al., 2018). Answers to this question were converted to a binary variable taking the value 1, if the respondent’s answer was greater than the median response of all the other participants of their country, and 0 otherwise.

Cognitive reflection was captured using the three-item Cognitive Reflection Test (CRT) (Frederick, 2005). The CRT evaluates a person’s propensity to ignore an alluring incorrect answer in favour of the correct answer, distinguishing those that overcome impulse, gut-instinct answers, and afford themselves a higher degree of reflection. We construct a binary indicator that takes value 1 for respondents who answered none of the CRT questions correctly, and 0 otherwise.

Behavioural traits are hypothesised to play an important role in shaping intentions to freelance. Higher risk aversion is expected to reduce willingness to engage in income-uncertain work arrangements, while an internal locus of control may increase intentions to freelance. Lower cognitive reflection, associated with greater impulsivity and impatience, may be linked to higher intentions to freelance, and positive reciprocity may increase interest in freelance work given its possible reliance on trust-based, relational contracting.

Socio-demographic characteristics The last sections of the survey collected information on the respondents socio-demographic profiles. We specifically collected information on the respondents’ age, gender, whether they are originally

migrants,¹¹ the number of years they spent in education¹² and whether they have children.¹³ The survey data also includes information on the respondents' broad income category, which was standardised to low, medium or high income relative to their country of residence.¹⁴ Several of these characteristics have been documented as key determinants of individual decision to transition to entrepreneurship (Simoes et al., 2016; Evans and Leighton, 1989).

Socio-demographic characteristics are also expected to be associated with intentions to freelance. Younger individuals, those without dependent children, and individuals with lower opportunity costs of leaving salaried employment may be more likely to consider freelancing. At the same time, higher education and income may either facilitate freelancing through greater human capital or discourage it by increasing the opportunity cost of leaving stable employment, making the expected relationship theoretically ambiguous.

3 Data

3.1 Sample Selection

Our original dataset contains responses from 18,019 subjects across 16 countries. Starting from this survey data, we exclude individuals based on several criteria. Notably, those identifying as students or retirees or already in self-employed at the time the survey was administered are excluded, even if working part-time. This is because we are solely interested in individuals currently working and not already engaged in self-employed activities. We also drop members of the armed

¹¹The specific question asked whether the individual was born in their current country of residence.

¹²Participants were asked to exclude pre-school, but include post-secondary education and training in their count.

¹³The survey did not collect the age of the children, or whether they are still legal dependants on the participant's income.

¹⁴This income includes salaries, pensions, benefits and allowances. Appropriate income categories were chosen by the survey provider for each country, as was the most common reference period (i.e. week, month or year).

forces. Additionally, we exclude responses from individuals who did not answer the occupation question or declared to be civil servants but provided no other indication of their occupation. Some respondents provided answers to the occupation questions that could not be classified into ISCO-08 occupation codes. We retain these observations and assign them a separate occupation code.¹⁵ Data is jointly available on employment contract type, job history (number of past jobs), and income category. Observations with missing income are therefore excluded only as part of this broader intersection of required covariates, rather than being dropped solely due to missing income. Eliminating observations as per the above criteria leaves us with a sample of 13,483 individuals across the 16 countries.

3.2 Data Description

Table A.7 describes our data. The first row of the Table shows that around 18% of people in our sample disclosed an interest to engage in freelance, on-demand self-employment within the year.

Turning to background characteristics, around 46% of the respondents in our final sample are female. On average, individuals have completed between secondary and tertiary education, equivalent to around 13.6 years of education and have around one child each. A significant share (i.e. 15%) of workers in our sample are of foreign origins.

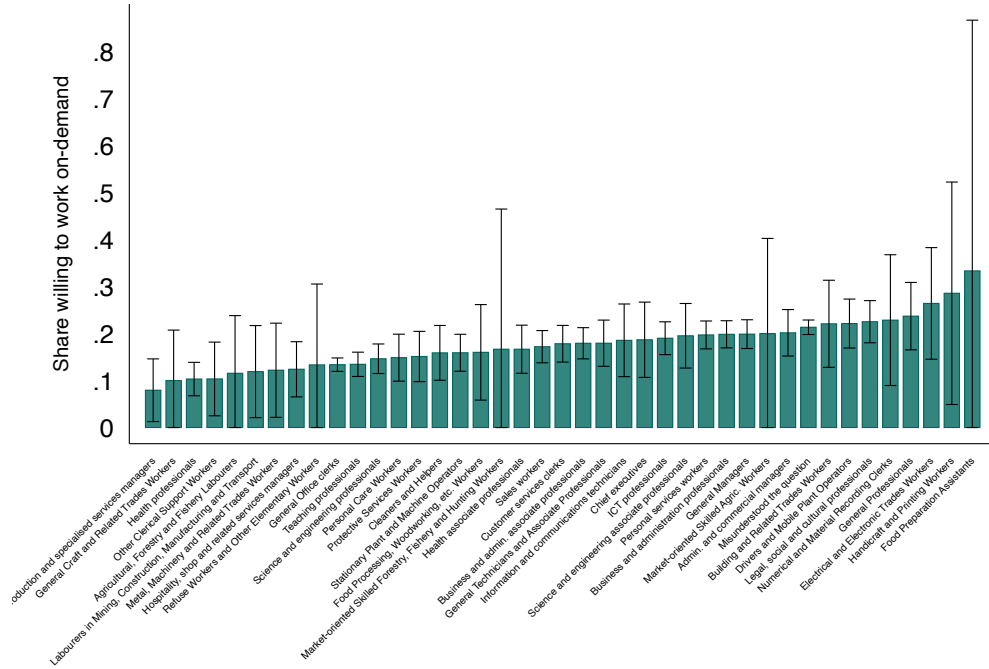
Looking at variables related to the respondents' work situation, in line with previous literature, we divide respondents between atypical workers and those who are employees with permanent contracts. We have 27% of individuals in our dataset have atypical work arrangements, 60% in private sector jobs and

¹⁵Answers to the occupation question were provided as a free text box. After translating the text answers to English, manual assignment of entries to standard occupational codes was performed. Often-appearing words were matched to detailed descriptions of standard ISCO occupational categories available at <https://www.ilo.org/public/english/bureau/stat/isco/isco08/>. Whenever the answer text was too short for the match to be successful, respondents are assigned a separate occupation code of 99. Table A.8 in Appendix B shows that our main results are robust to dropping these observations.

the sample on average working 35 hours per week. When asked to classify the task content of their job, half of respondents reported performing predominantly manual tasks as opposed to knowledge-based tasks, and 37% reported carrying out creative rather than routine tasks.

In the last 4 rows of Table A.7, we report summary statistics for measures of risk aversion, locus of control, cognitive reflection test (CRT) scores and positive reciprocity. We observe that 54% of our sample are risk averse, 61% have an internal locus of control, 53% classify as having a low CRT (by which they failed all CRT questions) and half the population can be characterised as having positive reciprocity.

Figure 1: Intentions to freelance by occupations



Notes: Share of respondents who intend to freelance, separately by occupation. Black caps represent 95% confidence intervals.

We also explore intentions to engage in freelance, on-demand self-employment

by occupation in Figure 1. We observe, in line with Bates (1995); Lofstrom et al. (2014); Chandler and Hanks (1998); Woronkiewicz and Noonan (2019) that intentions to freelance vary with workers' current occupation. In particular, around a third of workers employed as Food Preparation Assistants, Handicraft and Printing workers, and Electrical & Electronics Trades Workers report intentions to go freelance. This is in opposition to Health Professionals, undefined Craft Workers, and Production & Specialised Services Managers, of whom less than a tenth state interest in freelancing within the year.¹⁶

Figure 2 reports and highlights significant differences in intentions to freelance across our 16 sampled countries. Spain, Germany Finland, UK and Ireland are countries with no more than 10% of their respondents declaring interest in freelance, on-demand self-employment. Conversely, Malaysia and UAE have the highest population interest with just over 40% of their respondents interested in on-demand work. These intentions levels bare a slight positive relationship with the country's labour force proportion already in self-employment, as charted in Figure 2.¹⁷ Reasons for country-level self-employment heterogeneity such as fiscal motives or the national share of workers in the public sector Torrini (2005) could also be relevant for interest in freelancing. However, this figure does highlight some data points unexplained by national self-employment rates, such as US, Japan, Hong Kong and UAE, that seem adrift of the positive linear relationship. The weak association between national self-employment rates and our measure of intentions to enter freelance, on-demand self-employment highlights the limitations of aggregate self-employment measures and strengthens the case for distinguishing between different forms of self-employment in national statistics. We note a possible second pattern with our national data. In line with Wyrwich (2013) we note a possible positive association for lower intentions to

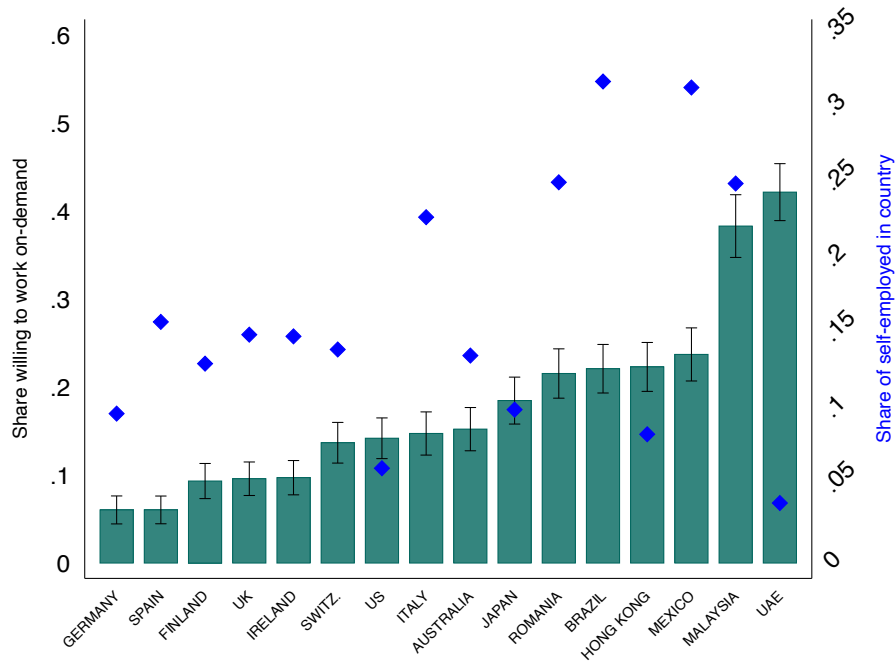
¹⁶Likely due to the low number of individuals in each of these occupation groups, we have large standard errors and cannot draw any statistical inference on these observations.

¹⁷As the survey data was collected in early 2019, we use ILO Stat data from the previous year, 2018, as opposed to more recent data.

freelance for countries that have more of a socialist, redistribution and welfare state culture.

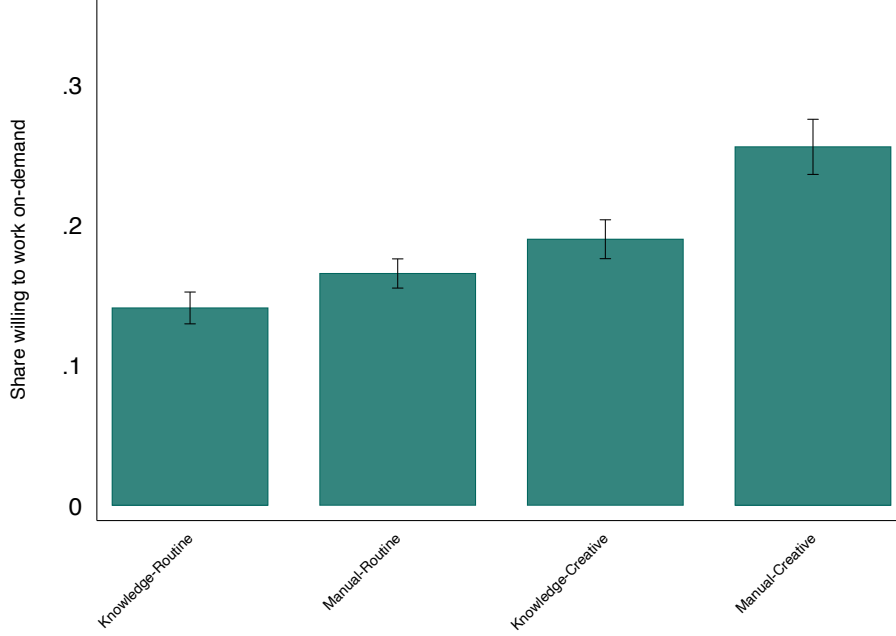
Figure 3 compares workers' interest in on-demand work based on the nature of their work task-type. We note a roughly 10 percentage point difference between the task-type combination that is most likely (manual-creative) and the least likely (knowledge-routine) to consider freelancing. We note that participants in creative roles, as opposed to routine roles, are those most likely to report an interest in freelancing.

Figure 2: Intentions to freelance by country



Notes: Share of respondents reporting positive intentions to freelance (left-hand y-axis), separately by country. Black caps represent 95% confidence intervals. National self-employment rates (right-hand y-axis) from 2018 ILO-Stat presented in blue diamonds.

Figure 3: Intentions to freelance by task type



Notes: Share of respondents reporting positive intentions to freelance, separately by task type. Black caps represent 95% confidence intervals.

4 Determinants of intentions to freelance

4.1 Empirical specification

The core of our empirical analysis relies on the estimation of a logit model with country and occupation fixed effects to uncover the determinants of freelancing intentions. The regression specification is:

$$\Pr(Freelance_{i,c,j} = 1) = \Lambda(\alpha_0 + \alpha_1 Socio_{i,c,j} + \alpha_2 Work_{i,c,j} + \alpha_3 Traits_{i,c,j} + u_c + u_j + \epsilon_{i,c,j}) \quad (1)$$

where $Freelance_{i,c,j}$, our dependent variable, indicates a respondent's intention to quit their main job and engage in freelance, on-demand self-employment within

the next 12 months, where i indicates the respondent, c her country of residence and j her occupation. $Freelance_{i,c,j}$ is equal to 1 if the individual declares she wants to freelance, and 0 otherwise.¹⁸

The term u_c stands for the country fixed effects which are included in all our regressions to remove the impact on freelancing intentions of fixed country characteristics that are potentially correlated with our main variables of interest. u_j refers to a set of occupation dummies.

We include three sets of variables in our regressions, $Socio_{i,c,j}$, $Work_{i,c,j}$ and $Traits_{i,c,j}$. The first set includes a battery of variables identifying individual characteristics such as age, gender, income categories, presence of children in the family, educational attainment and migration background. These variables are included in our model as they might independently affect one’s need to engage in on-demand work. These variables have been previously found to influence entry decisions into other forms of self-employment, such as entrepreneurship due to financial, social or opportunity barriers (Van der Sluis et al., 2008; Poschke, 2013; Koellinger et al., 2013; Wagner, 2007; Clark and Drinkwater, 2000; Borjas, 1986). Gender, for example, not only affects the reason for entering entrepreneurship, but also the propensity and job sector chosen (Williams, 2012; Stefanović and Stošić, 2012; Georgellis and Wall, 2005; Bates, 1995). Age and income are described with nonlinear relationships to entrepreneurship (Caliendo et al., 2014; Parker, 2004; Evans and Leighton, 1989; Blanchflower and Oswald, 1998; Fairlie and Krashinsky, 2012).

Our model also includes a suite of variables capturing job and employment characteristics ($Work_{i,c,j}$). We include a dummy identifying whether the respondent works in the private or public sector, the number of hours worked per week and whether the respondent’s work arrangement is atypical in nature. We also

¹⁸We also include regression analysis with an ordered logit, where the dependent varies from 1 “Very Unlikely”, 2 “Somewhat Unlikely”, 3 “Somewhat Likely” and 4 “Very Likely” to consider freelancing in the next 12 months (See Appendix Table A.5). We find consistent results with our binary dependent variable results presented in the next Section 4.2.

evaluate the respondent’s employment history, i.e. jobs held in the past. Importantly, we also include indicators for self-perceived task-types (Acemoglu and Autor, 2011), following previous literature indicating job and industry associations with self-employment (Blanchflower and Meyer, 1994; Bates, 1995; Woronkiewicz and Noonan, 2019). Additionally, we control for an indicator that takes value 1 if the worker perceives the likelihood of losing her job in the next 12 months as moderately or very high.

Finally, with $Traits_{i,c,j}$, we include behavioural traits such as risk aversion, locus of control, a new variable not typically discussed but relevant for labour choices, the cognitive reflection test (CRT) and positive reciprocity which, as previously shown in the literature, are important determinants of self-employment decisions (Mueller and Thomas, 2001; Caliendo et al., 2009; Brockhaus Sr, 1980; Beugelsdijk and Noorderhaven, 2005; Ekelund et al., 2005; Kihlstrom and Laffont, 1979; Stewart Jr and Roth, 2001).

Since we specify a logit model to represent choice behaviour, $\Lambda(\cdot)$ refers to the cumulative distribution function of the logistic distribution.

4.2 Main Results

We report our main model regression results in Table 1.

Table 1: Determinants of intentions to freelance

	(1)	(2)	(3)	(4)
Age (years)	-0.0676*** (0.0194)			-0.0667*** (0.0210)
Age (squared)	0.0558** (0.0244)			0.0556** (0.0259)
Female	-0.2961*** (0.0608)			-0.2633*** (0.0545)
Education	-0.2276* (0.1184)			-0.1812 (0.1118)
Education (squared)	0.0428 (0.0332)			0.0382 (0.0314)
Number of children	0.0174 (0.0226)			0.0108 (0.0232)
Migrant	0.0860 (0.1139)			0.0232 (0.0993)
Medium income	-0.1687* (0.0864)			-0.0543 (0.0927)
High income	-0.2968*** (0.0639)			-0.1097 (0.0705)
Number of past jobs		0.0016 (0.0069)		0.0184*** (0.0071)
Manual-Creative		0.4446*** (0.1000)		0.3846*** (0.0954)
Knowledge-Routine		-0.1441** (0.0564)		-0.0743 (0.0597)
Knowledge-Creative		0.1806** (0.0729)		0.2140*** (0.0727)
Atypical work arrangement		0.2328*** (0.0886)		0.1848** (0.0843)
Private sector		0.2675*** (0.0771)		0.2450*** (0.0666)
Hours worked		-0.0374* (0.0192)		-0.0313 (0.0220)
Likely job loss		1.3179*** (0.1101)		1.3261*** (0.1084)
Risk averse			-0.2694*** (0.0500)	-0.2362*** (0.0412)
Internal locus			0.0233 (0.0528)	0.1351* (0.0707)
Low CRT			0.2216*** (0.0516)	0.1627*** (0.0441)
Positive reciprocity			0.2845*** (0.0834)	0.2564*** (0.0726)
Constant	0.5168 (0.3611)	-2.2824*** (0.2077)	-1.8872*** (0.0861)	-0.5362 (0.4587)
Observations	13457	13457	13457	13457
Pseudo R^2	0.0851	0.1390	0.0783	0.1553
BIC	11600	10960	11637	10755
Country F.E.	yes	yes	yes	yes
Occupation F.E.	no	yes	no	yes

Notes: Logit regressions. Standard errors clustered at the country level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is a binary indicator taking value of one if the respondent reports being willing to engage in freelance work, and zero otherwise. Occupation dummies refer to the ISCO 2-digit classification of occupations. The reference category for task types is 'Manual-Routine' tasks.

We start, in Column (1), with a baseline logit specification that only accounts for socio-demographic factors, all the while controlling for country fixed effects.¹⁹ The results suggest a significant negative, and non-linear, relationship between age and stated intentions to go freelance. This is a novel finding that seems to oppose the entrepreneurship literature (Georgellis et al., 2005; Caliendo et al., 2014).²⁰ We note however that without data on entry into freelancing by age, we cannot distinguish genuine age-related preference from survivorship bias, as the older participants in our dataset will have necessarily foregone more opportunities to freelance. The survivorship effect does not explain the u-shape of age and freelance interest.²¹

Consistent with past work, women are 0.74 times $e^{-0.2961}$ less likely to state an interest in freelancing (Bates, 1995; Vejsiu, 2011; Koellinger et al., 2013; Clain, 2000; Clark, 1997). In opposition to the entrepreneurship literature and the liquidity constraint hypothesis, higher levels of income are negatively related to intentions to freelance, on-demand self-employment (Evans and Leighton, 1989, 1990; Bates, 1995; Blanchflower and Oswald, 1998; Fairlie and Krashinsky, 2012).²²

In column (2), we focus on the individual’s current work situation, controlling for both country and occupation fixed effects. We note, of the task types (Acemoglu and Autor, 2011), those in creative tasks, whether they be creative-manual

¹⁹Our results are robust to probit estimation procedures. The estimates are available from the authors upon request. The results are also robust to the sub-sample only ‘very likely’ to freelance, see Fig A.10 and ordered logit see A.5.

²⁰Although, Hintermaier and Steinberger (2005) do suggest that age would be negatively associated with self-employment rates due to lower physical and mental capacity for long work weeks and stressful work, which is not uncommon in freelancing work such as gig-work (Graham et al., 2017; Wood et al., 2019; Woodcock and Graham, 2019) but this does not explain the non-linearity.

²¹We also plot the intentions to freelance by age in Figure A.1 to further highlight the non-linearity. In Table A.1 we note that the age effects are significant only with female participants, the increased interest among older women might be linked to higher interest in future planning among this cohort (Clark et al., 2019).

²²Freelance, on-demand work typically features lower entry barriers than entrepreneurship and may therefore serve as a short-term labour market option for necessity-driven workers (Block and Sandner, 2009; Poschke, 2013).

or creative-knowledge types, are more likely to have stated interest in freelancing than those in manual-routine jobs, with 1.56 and 1.19 increased likelihood respectively. Amongst other job characteristics, a person in an atypical work arrangement is more likely, than those who do not, to be interested in freelancing. This reflects suggestions that those with permanent contracts have greater opportunity costs by transitioning to on-demand work, and so less likely to do so (Georgellis et al., 2005). Workers who believe they will lose their job within the next 12 months are significantly (3.74 times) more likely to be interested in freelance, on-demand self-employment compared to those who expect continuous salaried work. This is the strongest predictor in our main results table.

For column (3), our model estimates the relevance of behavioural variables. As with previous research, more risk averse individuals are generally less likely to express intentions to freelance (Caliendo et al., 2014; Ekelund et al., 2005; Stewart Jr and Roth, 2001; Kihlstrom and Laffont, 1979). Those with a low CRT exhibit higher intentions to freelance. While this is a novel finding with this metric, it is consistent with evidence showing that entrepreneurs display higher functional impulsivity, a trait negatively correlated with CRT scores (Lawrence et al., 2008). In addition, individuals with higher levels of positive reciprocity are more likely to intend to freelance, in line with literature emphasising the role of reciprocity in building professional networks and accessing work opportunities in self-employment (Baron and Markman, 2000; Caliendo et al., 2012, 2014). The magnitudes of these effects are similar, with low CRT and positive reciprocity increasing the likelihood of intending to freelance by approximately 1.25 and 1.33, respectively. By contrast, we find no significant association with internal locus of control, suggesting that behavioural traits relevant for entrepreneurial entry may not fully translate to intentions to engage in freelance, on-demand self-employment, consistent with distinctions between opportunity- and necessity-driven transitions (Simoes et al., 2016; Block and Sandner, 2009).

We combine in column (4) Table 1 all sets of variables to present a more accurate picture of the importance of each factor, and test their robustness in a more complete model. Consistent with column (1) age keeps its u-shaped relationship with interest in freelance, on-demand self-employment, while income loses its significance. From column (2), we note that all job characteristics, apart from hours typically worked in a week and the knowledge-routine task type, are robust to the more complete regression model, as are our behavioural variables from column (3).

5 Robustness

We run a number of tests to ascertain sensitivity of our results, for example whether our discovered key predictors are stable across our 16 countries. We also address potential omitted variables biases, notably for current working and health conditions, and explore any differences in explanatory power of our main model specification.

Country-level The inclusion of country dummies in our previous regressions could conceal heterogeneity on our key variables due to jurisdictional specific factors, including institutional ones.²³ We therefore augment the baseline specification with a set of country-level variables capturing institutional, economic, and labour market characteristics. The results are presented in Table A.12.²⁴ Specifically, we include the Global Rights Index (ITUC), which captures cross-country variation in workers' rights (column 1), the share of private sector employment (column 2), the level of self-employment (column 3), and the growth in self-

²³We additionally verify the robustness of our findings by applying a leave-one-group-out approach, whereby each regression is re-estimated after excluding one country at a time (Appendix Table A.6). We find no evidence that any single country consistently influences the magnitude or sign of the coefficient estimates.

²⁴All national-level variables refer to 2018 to align with the timing of the survey (February–March 2019).

employment between 2008 and 2018 (column 4).²⁵ These variables have been discussed in prior work as relevant correlates of self-employment and labour market structure (Grubb and Wells, 1993; Acs et al., 1994; Torrini, 2005). Across these alternative specifications, our core individual-level predictors, including age, education, income, task type, and risk aversion, remain consistent with the baseline results reported in Table 1. By contrast, the coefficients on gender and low cognitive reflection lose statistical significance, while internal locus of control becomes significant in some specifications.

Job prospects and satisfaction Our data do not include direct measures of job satisfaction or information on the specific type of freelance activity respondents intend to pursue. As a result, intentions to freelance may partly reflect dissatisfaction with current employment or self-selection related to career prospects rather than preferences for independent work per se. While we cannot address these concerns directly, we conduct robustness checks that control for respondents’ perceived autonomy in daily work activities and their influence over organisational decisions, which serve as proxies for job satisfaction (Table 2).²⁶

The results indicate that greater perceived autonomy and influence are positively associated with intentions to engage in freelance, on-demand self-employment. When additional job and health controls are included, influence over organisational decisions remains a robust predictor, while autonomy over daily tasks loses significance.²⁷

²⁵Workers’ rights data are sourced from the ILO/ITUC, while self-employment measures are obtained from the World Bank.

²⁶Locus of control, controlled for in our main regressions (Table 1), is also associated with perceived job satisfaction (Ng et al., 2006; Judge and Bono, 2001). Locus of control has a significant but weak Pearson correlation at the 1% level with both influence on one’s own company (as a continuous variable) and likelihood to lose job, 0.15 and -0.08 common variance, respectively.

²⁷In the Appendix, Table A.4 presents all the variable results of the full set of estimates.

Openness to experience Some personality traits are also omitted from our analysis due to data availability. In particular, we believe that openness to experience, would be relevant to our analysis. Recent literature has in fact shown that this trait positively affects subjective career success (Van den Born and Van Witteloostuijn, 2013) and job mobility (Nieß and Zacher, 2015). Schmitt-Rodermund (2004) and Obschonka et al. (2010) have shown that individuals with a stronger entrepreneurial personality exhibit high openness. Equally, Caliendo et al. (2014) identify openness to new experience as a strong predictor of entry into entrepreneurship. To ascertain the possible bias arising from the omission of openness to experience, we control for intentions to leave one’s job voluntarily, as a proxy for openness to experience (see Table 2 column (3)).²⁸

Controlling for workers’ intentions to leave their job voluntarily, as proxy for openness to new experiences, has the biggest positive contribution on our baseline model’s predictive power, adding 10 percentage points to the Pseudo- R^2 . Furthermore, as shown in column (3), intentions to leave one’s job voluntarily are highly predictive of freelancing intentions.

Health condition Previous reviews (Simoes et al., 2016) have suggested that workers’ health conditions affect entrepreneurial entry. Yet evidence in this respect remains mixed. Some studies find a positive association between poor health and the business ownership (Borjas, 1986; Zissimopoulos and Karoly, 2007; Pagán, 2009; Jones and Latreille, 2011), others find opposite results (Taylor, 2001; Parker and Rougier, 2007; Cahill et al., 2013). In our case, it can be argued that poor health mediates intentions to engage in freelance, on-demand self-employment. Such work arrangements may be perceived as offering greater flex-

²⁸It is worth noting that intentions to engage in freelance, on-demand self-employment are strongly correlated with intentions to leave one’s job voluntarily (correlation coefficient of 63%). This relationship is mechanically expected to some extent, as leaving salaried employment is a necessary condition for transitioning into freelance, on-demand work. It is therefore possible that some respondents interpreted the question on voluntary job leaving more broadly, as indicative of any job change, rather than a specific transition into self-employment, which would naturally generate a high degree of overlap between the two measures.

ibility to accommodate health limitations or disabilities alongside work responsibilities. To account for this possibility, we include a control for respondents' self-reported health status in column (4) of Table 2.²⁹ We find that self-perceived health status is not significantly associated with intentions to engage in freelance, on-demand self-employment.

Overall, in column (5), we note that including all the job and health conditions simultaneously in our regressions improves upon the predictive power of our main results model, increasing our Pseudo- R^2 from 15.38% to 23.71%.

Table 2: Robustness checks

	(1)	(2)	(3)	(4)	(5)
Have influence on own activity	0.2544*** (0.0787)				0.0453 (0.0628)
Have influence on firm		0.3685*** (0.0791)			0.4203*** (0.0811)
Likely to leave job			1.7138*** (0.0983)		1.7379*** (0.1058)
Better than avg. health				0.0665 (0.0503)	0.0529 (0.0496)
Constant	-0.7878 (0.5293)	-0.9959* (0.5427)	-2.2771*** (0.5285)	-0.5698 (0.4721)	-2.7866*** (0.6352)
Observations	11970	12625	13457	13457	11758
Pseudo R^2	0.1538	0.1585	0.2339	0.1554	0.2371
BIC	9541	9803	9768	10754	8390
Country F.E.	yes	yes	yes	yes	yes
Occupation F.E.	yes	yes	yes	yes	yes

Notes: Logit regressions including all variables from Table 1 column (4). Standard errors clustered at the country level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is a binary indicator taking value of one if the respondent reports being willing to engage in freelance work, and zero otherwise. Occupation dummies refer to the ISCO 2-digit classification of occupations.

6 Heterogeneity

We also investigate interaction effects, notably studying the moderating effects of age and income, age and risk tolerance, income and risk tolerance, income and

²⁹The health variable is a binary indicator equal to one if respondents report being healthier than others of the same age, and zero otherwise.

fear of job loss, and gender with risk tolerance.

Age and income Previous studies have suggested and empirically confirmed the non-linear impact of age on entrepreneurial decisions (Hintermaier and Steinberger, 2005; Caliendo et al., 2014), as we observe in our own results in Table 1. Age is also responsible for within country heterogeneity in economic preferences (Falk et al., 2018). We study some of these possible interactions in Figure 4a. We note an overall trend of decreasing interest in freelancing by income, while holding age fixed, as found in our main results Table 1 and Appendix Table A.2. We note that younger workers (below 44 years of age) who currently earn low income have the highest intentions to engage in on-demand independent work. Conversely, older workers currently earning high income have the lowest interest in freelancing.³⁰

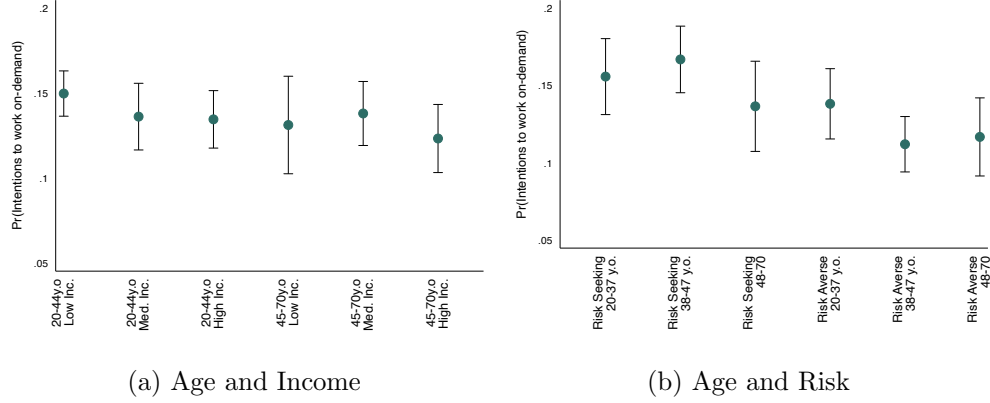
Age and risk tolerance It can be hypothesised that with age, risk prospects might vary strongly. For example, young adults just joining the labour market might not be too concerned with career stability and more likely to experiment with job options and consider freelance, on-demand self-employment. On the other hand, adults wishing to start a family might be less likely to seize financially risky and less stable work opportunities. Additionally, in older years, with more work experience, social and financial capital, freelance, on-demand self-employment might become a more interesting prospect. We study these possible interactions in Figure 4b.³¹ It is noted that risk preferences transcend age. Risk seeking individuals who are 20-47 and 48-70 years of age have roughly the same level of interest in freelancing, approximately 17.1%. Risk averse participants in the two older age groups (38-47 and 48-70) have very similar disinterest in

³⁰We note that the 95% confidence intervals are large for certain age income interactions, leaving few of our results in this panel significant.

³¹Larger age brackets than Figure 4b were used to reduce the heterogeneous and low power of some age-income interactions.

freelancing, approximately 12.3%.

Figure 4: Predicted probability of freelancing intentions by income level, age and risk



Notes: Predicted probability of expressing freelancing intentions for the six possible combinations of age and income categories (panel a), and age and risk preference (panel b). Respondents are classified as risk averse compared to their national median level. The dots represent marginal effects from the interaction variable, and the bars show 95% confidence intervals. All regressions include the full set of control as per Table 1 column (4).

Income and risk tolerance One's financial situation is described as an important determinant of self-employment choices. Easy access to assets and income might help minimise constraints to entry into self-employment - especially opportunities entailing high upfront costs - as well as offer a safety net for unexpected adverse events (Hundley, 2006; Nykvist, 2008). Conversely, a lack of income will lead to general occupation choices by necessity, as opposed to opportunity (Block and Sandner, 2009; Poschke, 2013). We ponder how risk might interact with income levels and interest in freelancing. On one hand, risk-averse individuals in the upper income brackets might be more likely to see financial risks associated with freelancing as too high to consider the transition away from stable positions. On the other hand, those considering transitions by necessity – more likely to be low income individuals – will have less choice and may have to consider freelancing, regardless of one's risk preferences. The results presented in Figure 5a are in

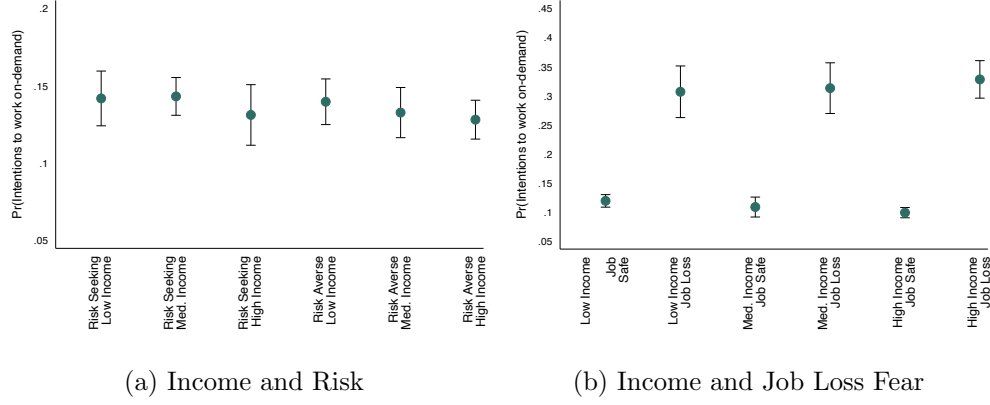
line with these hypotheses. It is shown that risk tolerance moderates intentions to freelance among the high income earners, with high-income risk-seeking individuals significantly more likely to have stated interest in freelancing than their high-income risk-averse counterparts. We also observe, however, that among the low and medium income earners, risk preference does not significantly moderate interest in freelancing.

Income and fear of job loss In our main results from Column 4 Table 1 we observed a negative relationship between income and intentions to freelance, particularly between the highest and lowest income groups. Perhaps freelance, on-demand self-employment is sought, not as an opportunity to step up and increase future prospects, but by necessity to find a new job to meet life subsistence needs (Scarelli and Margolis, 2023). We test this idea in Figure 5b by interacting income with fear of job loss. We note a clear and significant increase in stated interest in freelance, on-demand work when fearing job loss, this is true across all income groups. Amongst those fearing loss of their job in the near future, the negative income relationship does not hold. However, for those who believe they have short-term job security, the negative income relationship is still significant at the 5% level between low and high-income earners.

Gender and risk tolerance Observable discrepancies in male and female representation in self-employment are potentially linked to omitted variables. Risk tolerance is commonly discussed as a potential reason for this, with women relatively more risk averse than men and consequently less inclined to freelance than men (Dohmen et al., 2010; Caliendo et al., 2009; Croson and Gneezy, 2009).³² Figure 6 shows that risk preferences do moderate the interest in freelancing, risk-seeking women (men) are significantly more likely to consider freelancing than

³²Other variables of influence include, job satisfaction and desire to change employment (Ng et al., 2006; Judge and Bono, 2001), and representation differences in sectors, which have varying transition to self-employment opportunities (Bates, 1995; Vejsiu, 2011)

Figure 5: Predicted probability of freelancing intentions by income level, risk and fear of job loss



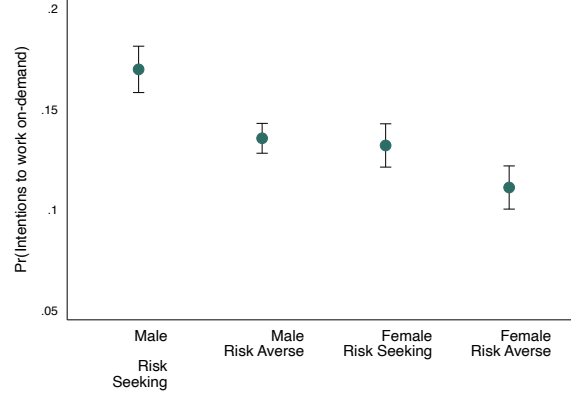
Notes: Predicted probability of expressing freelancing intentions for the six possible combinations of income and risk categories (panel a) and 6 combinations of income and fear of job loss (panel b). Respondents are classified as risk averse compared to their national median level. They are classified as fearing job loss if they stated a high likelihood of losing their job within the next 12 months. The dots represent marginal effects from the interaction variable, and the bars show 95% confidence intervals. All regressions include the full set of control as per Table 1 column (4).

risk-averse women (men). But, between women and men, risk-seeking men are still significantly more likely to consider on-demand work than risk-seeking women (14.1% and 11.8%, respectively).

Risk aversion and freelance intention at the national level In our main results (Table 1) and the previous heterogeneity analysis (Figures 4b, 5a and 6), we noted the different circumstances in which risk preferences may influence stated intentions to freelance at the individual level. In all of these cases, we considered an individual's risk aversion as relative to the median of their country, while controlling for country dummies in our regressions. However, we know that absolute risk aversion does vary across countries Falk et al. (2018), and so in Table A.11 we consider the country-level variation in influence of risk preferences.

Out of our 16 country-level regressions, only four countries (Germany, Japan,

Figure 6: Predicted probability of freelancing intentions by gender and risk preference



Notes: Predicted probability of expressing freelancing intentions for the four possible combinations of gender and risk preferences. Respondents are classified as risk averse compared to their national median level. The dots represent marginal effects from the interaction variable, and the bars show 95% confidence intervals. All regressions include the full set of control as per Table 1 column (4).

Spain and Switzerland) present risk preferences as a significant predictor of intentions to freelance. There is seemingly heterogeneity in the importance of a typical predictor in the literature. This country-level difference should be explored to determine why risk preferences are pivotal in some countries and not others.

Years of education, used as a proxy for skill level, are employed to assess whether our results differ between individuals with education levels below and above their country-specific median. Table A.3 shows that most key estimates are broadly consistent across the two education groups. The main exception concerns private sector employment and atypical contracts, which are statistically significant only among individuals with above-median education levels.³³

³³Across both education groups, being older, female, or more risk averse is associated with lower intentions to freelance, while lower CRT scores, higher positive reciprocity, employment in creative tasks, and a higher perceived likelihood of job loss within the next 12 months are associated with greater intentions to freelance. Private sector employment exhibits the strongest heterogeneity by education level, being significant only among the above-median group. Table A.9 further shows that including country and task-type fixed effects does not alter these results.

7 Discussion and conclusion

Own-account, project-based (on-demand) work has outgrown its former status as a niche segment of the labour market. The rise of complex portfolio careers, which eschew traditional career paths, has burgeoned in parallel with the widespread adoption of digital technologies. In this context, workers can choose how to craft their complex work portfolios, selecting the tasks across multiple employers (Burke and Cowling, 2020b), while firms benefit from the ability to outsource discrete tasks with reduced labour market frictions and lower operational costs.

At the same time, these activities are associated with several well-documented psychological, welfare costs and financial risks (Graham et al., 2017; Wood, 2016; Wood et al., 2019; Woodcock and Graham, 2019; MacDonald and Giazitzoglu, 2019; Hyytinen et al., 2013; Hamilton, 2000). Depending on the institutional and legal context, workers engaged in this form of self-employment may face reduced access to social protection relative to standard employment, or be expected to self-insure against risks typically covered in salaried jobs (Prassl and Risak, 2015; Prassl, 2018). The extent and nature of these challenges vary substantially across countries, reflecting differences in labour law, social insurance systems, and regulatory approaches.

Against this backdrop, understanding the key determinants of workers' decisions to transition from salaried employment into freelance, on-demand self-employment is increasingly important. Using a novel cross-country survey of workers currently carrying out different tasks and occupations, we provide descriptive evidence on intentions to leave salaried employment to engage in this specific mode of self-employment across 16 countries. Our results identify which groups of workers are more likely to consider such transitions and reinforce the importance of clearly distinguishing between different forms of self-employment in both research and policy design (Kässi and Lehdonvirta, 2018; Audretsch et al., 2015).

Several of our findings point to concerning pathways into freelance, on-demand self-employment that may contribute to inequality. Low-income individuals are more likely to consider transitioning into this form of work than higher-income workers, and while risk preferences moderate intentions among higher-income individuals, this pattern does not hold for lower-income earners. Workers in atypical employment arrangements, those performing creative work tasks, individuals with lower cognitive reflection scores, and those expressing fear of job loss also exhibit higher intentions to enter freelance, on-demand self-employment. Younger workers are more likely than older individuals to consider such transitions, and higher intentions are observed among women in atypical contracts, private-sector employees, and those with lower levels of educational attainment. Taken together, these results suggest that there is a portion of the population drawn into freelance, on-demand self-employment less by opportunity or preference and more by necessity, financial pressure, or a limited appreciation of the longer-term risks associated with these work arrangements.

A key limitation of this study is that we analyse workers' intentions to engage in freelance, on-demand self-employment rather than realised labour market transitions. While stated intentions do not necessarily translate into observed behaviour, they provide valuable information about workers' preferences, expectations, and perceived constraints at the margin. This distinction is particularly relevant in the context of freelance, on-demand work, where entry barriers are relatively low and transitions may occur rapidly, intermittently, or possibly in combination with other forms of employment. Moreover, given the scarcity of datasets that clearly distinguish freelance, on-demand self-employment from other forms of self-employment across countries, analysing intentions represents an important step forward in identifying who is at risk of transitioning into this form of work and why. That said, intentions cannot capture the full dynamics of actual labour market trajectories, including persistence in freelance work, earn-

ings paths, or subsequent returns to salaried employment. Future research linking stated intentions to realised behaviour using longitudinal or administrative data would therefore be essential.

Going forward, our findings point to several directions for future research. First, informational interventions studies should be considered for groups that are most likely to consider freelancing. These studies should test whether misinformation gaps, impulsiveness and narrow bracketing -ignoring longer term sacrifices and risks - cause an increased interest in freelance, on-demand self-employment.

Second, we noted that several vulnerable cohorts stated higher intentions to freelance, likely in part by necessity to cover living needs. Policy makers might consider targeted financial support mechanisms that enhance workers' bargaining power and time for a salaried job position or provide stability during transitions. Such measures might include temporary income support, default enrolment in income protection schemes in case of job cessation for sectors or groups considering to freelance by necessity.

Third, to assist with targeted policy interventions and the development of labour legislation, more precise measurement and classification of on-demand work are essential for effective policy design and regulation (Prassl, 2018; Kässl and Lehdonvirta, 2018; Prassl and Risak, 2015; Audretsch et al., 2015). Improved data collection would facilitate clearer distinctions across forms of self-employment and facilitate the development of targeted regulatory responses such as minimum payment standards, access to parental leave benefits, or revenue-smoothing mechanisms aimed at mitigating income volatility among vulnerable freelancers.

Finally, future research should examine the longer-term labour market trajectories associated with freelance, on-demand self-employment, including transitions back into salaried work and potential segmentation or lock-in effects.

Overall, our results highlight the need for a coordinated effort across labour

law, public policy, and economic research to ensure that the continued expansion of freelance, on-demand self-employment does not exacerbate existing inequalities or create new forms of labour market vulnerability.

A Additional Descriptive Evidence

Table A.1: Gender and intentions to freelance

	Female Only	Male Only
Age (years)	-0.0865*** (0.0207)	-0.0435 (0.0307)
Age squared	0.0765*** (0.0254)	0.0310 (0.0374)
Educational attainment	-0.2937*** (0.1361)	-0.0625 (0.1401)
Education Squared	0.0692** (0.0425)	0.0048 (0.0361)
Number of children	0.0160 (0.0311)	-0.0080 (0.0307)
Migrant	0.0558 (0.1160)	-0.0132 (0.1673)
Medium income	-0.0698 (0.0721)	-0.0536 (0.1599)
High income	-0.1420 (0.0926)	-0.0814 (0.1378)
Number of past jobs	0.0142* (0.0074)	0.0244** (0.0117)
Manual-Creative	0.3350** (0.1417)	0.4494*** (0.1000)
Knowledge-Routine	-0.1524* (0.0868)	0.0327 (0.0886)
Knowledge-Creative	0.1778* (0.1024)	0.2754*** (0.0895)
Atypical work arrangement	0.2221** (0.0953)	0.1171 (0.1322)
Private sector	0.3304*** (0.1003)	0.1175* (0.0656)
Hours worked	0.0051 (0.0369)	-0.0698*** (0.0210)
Likely job loss	1.3612*** (0.1251)	1.2815*** (0.1128)
Risk averse	-0.2375*** (0.0535)	-0.2217*** (0.0687)
Internal Locus	0.1757*** (0.0630)	-0.0767 (0.1182)
Low CRT	0.1814*** (0.0502)	0.1384* (0.0715)
Positive Reciprocity	0.2927*** (0.0777)	0.2010* (0.1040)
Constant	-0.2135 (0.4766)	-1.2875* (0.6617)
Observations	7205	6207
Pseudo R^2	0.1667	0.1462
BIC	6079	4719
Country F.E.	yes	yes
Occupation F.E.	yes	yes

Notes: Logit regressions. Standard errors clustered at the country level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is a binary indicator taking value of one if the respondent reports being willing to engage in freelance work, and zero otherwise. Occupation dummies refer to the ISCO 2-digit classification of occupations. The reference category for task types is 'Manual-Routine' tasks.

Table A.2: Income and intentions to freelance

	Low Income	High Income
Age (years)	-0.0506 (0.0312)	-0.1138*** (0.0222)
Age squared	0.0359 (0.0378)	0.1074*** (0.0286)
Female	-0.2816*** (0.1056)	-0.1880** (0.0877)
Educational attainment	-0.2453 (0.1751)	-0.0480 (0.1465)
Education Squared	0.0441 (0.0549)	-0.0150 (0.0453)
Number of children	-0.0159 (0.0419)	0.0687* (0.0379)
Migrant	0.1646 (0.1211)	-0.0464 (0.2006)
Risk averse	-0.3046** (0.1363)	-0.2245*** (0.0550)
Internal Locus	0.0384 (0.0930)	0.2014*** (0.0715)
Low CRT	0.1681*** (0.0553)	0.0664 (0.0794)
Positive Reciprocity	0.1309 (0.1087)	0.2263*** (0.0775)
Number of past jobs	0.0072 (0.0148)	0.0227*** (0.0080)
Manual-Creative	0.3121** (0.1408)	0.4801*** (0.1586)
Knowledge-Routine	0.0240 (0.0861)	-0.1273 (0.1353)
Knowledge-Creative	0.2840** (0.1309)	0.2025** (0.0894)
Atypical work arrangement	0.2578*** (0.0826)	-0.1316 (0.1249)
Private sector	0.1104 (0.0936)	0.3334** (0.1440)
Hours worked	-0.0446 (0.0322)	-0.0371 (0.0413)
Likely to lose job	1.2098*** (0.1155)	1.5420*** (0.1004)
Constant	-0.5127 (0.7445)	0.2719 (0.5253)
Observations	4257	4944
Pseudo R^2	0.1593	0.1839
BIC	4038	4132
Country F.E.	yes	yes
Occupation F.E.	no	yes

Notes: Logit regressions. Standard errors clustered at the country level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is a binary indicator taking value of one if the respondent reports being very willing to engage in freelance work, and zero otherwise. Occupation dummies refer to the ISCO 2-digit classification of occupations. The reference category for task types is 'Manual-Routine' tasks. Only Low and high income groups are considered, not medium income.

Table A.3: Median education and intentions to freelance

	Lower Education	Higher Education
Age (years)	-0.0728** (0.0337)	-0.0641** (0.0292)
Age squared	0.0613 (0.0408)	0.0540 (0.0335)
Female	-0.3274*** (0.0822)	-0.2019*** (0.0778)
Number of children	-0.0229 (0.0421)	0.0414* (0.0238)
Migrant	0.1152 (0.0882)	-0.0453 (0.1681)
Medium income	-0.1572 (0.1420)	0.0466 (0.1024)
High income	-0.1490 (0.1241)	-0.0653 (0.1049)
Risk averse	-0.2340*** (0.0692)	-0.2306*** (0.0845)
Internal Locus	0.1305 (0.0987)	0.1482* (0.0757)
Low CRT	0.1446** (0.0720)	0.1774*** (0.0647)
Positive Reciprocity	0.2645*** (0.0724)	0.2498*** (0.0865)
Number of past jobs	0.0242* (0.0143)	0.0141** (0.0061)
Manual-Creative	0.4006*** (0.1277)	0.3711*** (0.1200)
Knowledge-Routine	-0.0230 (0.0971)	-0.0965 (0.0784)
Knowledge-Creative	0.3176*** (0.1039)	0.1616** (0.0806)
Atypical work arrangement	0.1935 (0.1290)	0.1893** (0.0875)
Private sector	0.0839 (0.0825)	0.3873*** (0.0780)
Hours worked	-0.0207 (0.0284)	-0.0392 (0.0448)
Likely to lose job	1.3319*** (0.1502)	1.3412*** (0.0941)
Constant	-0.1211 (0.6101)	-1.1523* (0.6539)
Observations	5859	7575
Pseudo R^2	0.1643	0.1552
BIC	4968	5830
Country F.E.	yes	yes
Occupation F.E.	yes	yes

Notes: Logit regressions. Standard errors clustered at the country level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is a binary indicator taking value of one if the respondent reports being very willing to engage in freelance work, and zero otherwise. The data is split into two groups, those with a lower level of education than their country's median, column (1) regression, and those with a above median level of education. Occupation dummies refer to the ISCO 2-digit classification of occupations. The reference category for task types is 'Manual-Routine' tasks. Low income group is the income variable reference.

Table A.4: Full Model with Job Conditions

	Long Model
Age (years)	-0.0191 (0.0240)
Age squared	0.0081 (0.0285)
Female	-0.2873*** (0.0596)
Educational attainment	-0.0434 (0.1181)
Education Squared	-0.0078 (0.0323)
Number of children	0.0199 (0.0226)
Migrant	0.0860 (0.1246)
Medium income	-0.0495 (0.0903)
High income	-0.1116 (0.0706)
Risk averse	-0.1305** (0.0539)
Internal Locus	0.1574* (0.0809)
Low CRT	0.2099*** (0.0557)
Positive reciprocity	0.2383*** (0.0761)
Number of past jobs	0.0141* (0.0075)
Manual-Creative	0.4149*** (0.0861)
Knowledge-Routine	-0.0643 (0.0524)
Knowledge-Creative	0.2251** (0.0933)
Atypical work arrangement	0.2462*** (0.0923)
Private sector	0.1890*** (0.0629)
Hours worked	-0.0523** (0.0237)
Likely to lose job	0.8339*** (0.1039)
Have influence on own activity	0.0453 (0.0628)
Have influence on firm	0.4203*** (0.0811)
Likely to leave job	1.7379*** (0.1058)
Better than avg. health	0.0529 (0.0496)
Constant	-2.7866*** (0.6352)
Observations	11758
Pseudo R^2	0.2371
BIC	8390
Country F.E.	yes
Occupation F.E.	yes

Notes: Logit regression. Standard errors clustered at the country level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is a binary indicator taking value of one if the respondent reports being willing to engage in freelance work, and zero otherwise. Occupation dummies refer to the ISCO 2-digit classification of occupations. The reference category for task types is 'Manual-Routine' tasks.

Table A.5: Ordered Logit on Freelancing - Main Results

	Main Results Model
Age (years)	-0.0390** (0.0182)
Age squared	0.0207 (0.0223)
Female	-0.3409*** (0.0544)
Number of children	-0.0185 (0.0147)
Migrant	0.0449 (0.0961)
Medium income	-0.0852 (0.0704)
High income	-0.0918* (0.0555)
Educational attainment	-0.1319 (0.1064)
Education Squared	0.0382 (0.0294)
Risk averse	-0.2126*** (0.0280)
Internal Locus	-0.0082 (0.0526)
Low CRT	0.0722* (0.0422)
Positive Reciprocity	0.2710*** (0.0652)
Number of past jobs	0.0166*** (0.0048)
Manual-Creative	0.3171*** (0.0861)
Knowledge-Routine	-0.0854 (0.0558)
Knowledge-Creative	0.2143*** (0.0665)
Atypical work arrangement	0.1790*** (0.0678)
Private sector	0.3593*** (0.0642)
Hours worked	-0.0442** (0.0217)
Likely to lose job	1.0990*** (0.0808)
Ordered Logit cut 1	-0.6040 (0.3730)
Ordered Logit cut2	0.8064** (0.4026)
Ordered Logit cut3	2.5418*** (0.4483)
Observations	13457
Pseudo R^2	0.0984
<i>BIC</i>	25850
Country F.E.	yes
Occupation F.E.	yes

Notes: Logit regressions. Standard errors clustered at the country level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is a categorical variable with values from 1 (very unlikely) to 4 (very likely) for stated freelancing intentions within the next 12 months. Occupation dummies refer to the ISCO 2-digit classification of occupations. The reference category for task types is 'Manual-Routine' tasks. Low income group is the income variable reference.

Table A.6: Leave-one country out estimation sensitivity

Coefficient	Min Estimate Dropped Country	Min Estimate Coefficient	Full Sample Regression	Max Estimate Coefficient	Max Estimate Dropped Country
Age (years)	-0.0765	UK	-0.0667	-0.0548	Japan
Age squared	0.0418	Japan	0.0556	0.0674	UK
Female	-0.3028	Malaysia	-0.2633	-0.2392	Switzerland
Educational attainment	-0.2438	UAE	-0.1812	-0.1344	Brazil
Education Squared	0.0237	Brazil	0.0382	0.0570	UAE
Number of children	-0.0003	Malaysia	0.0108	0.0233	Switzerland
Migrant	-0.0252	UK	0.0232	0.0984	UAE
Medium income	-0.1202	Brazil	-0.0543	-0.0088	Hong Kong
High income	-0.1413	Brazil	-0.1097	-0.0735	Hong Kong
Risk averse	-0.2613	Hong Kong	-0.2362	-0.2195	Germany
Internal Locus	0.0911	Japan	0.1351	0.1715	Italy
Low CRT	0.1389	Finland	0.1627	0.1847	US
Positive reciprocity	0.2210	Australia	0.2564	0.2944	Brazil
Number of past jobs	0.0160	Mexico	0.0184	0.0243	Hong Kong
Manual-Creative	0.3497	UK	0.3846	0.4475	UAE
Knowledge-Routine	-0.1197	Romania	-0.0743	-0.0529	Brazil
Knowledge-Creative	0.1758	Italy	0.2140	0.2642	Brazil
Atypical work arrangement	0.1423	Australia	0.1848	0.2411	US
Private sector	0.2149	Brazil	0.2450	0.2926	UAE
Hours worked	-0.0463	Mexico	-0.0313	-0.0178	UAE
Likely to lose job	1.2764	US	1.3261	1.3908	Brazil

Notes: Table summarising the original coefficient estimates, as in main results section, in addition to the minimum and maximum estimated coefficient for each variable when removing one country's worth of data at a time (LOGO = leave-one-group-out). The dropped country columns report the omitted country the minimum and maximum coefficients. Standard errors were clustered by country, occupation and country fixed effects were also included, as in the specification of the full regression in the Main Results section.

Table A.7: Descriptive statistics

	Mean	St. Dev.	N
Likely to freelance	0.177	0.382	13457
Age (years)	41.477	11.604	13457
Female	0.464	0.499	13457
Years of education	13.605	4.722	13457
Number of children	1.076	1.134	13457
Migrant	0.152	0.360	13457
Low income	0.317	0.465	13457
Med. income	0.314	0.464	13457
High income	0.369	0.482	13457
Abv. median past jobs	0.599	0.490	13457
Manual tasks	0.504	0.500	13457
Creative tasks	0.369	0.482	13457
Atypical work arrangement	0.270	0.444	13457
Private sector	0.601	0.490	13457
Hours worked	4.773	1.325	13457
Risk averse	0.541	0.498	13457
Internal locus	0.606	0.489	13457
Low CRT	0.537	0.499	13457
Likely to lose job	0.197	0.398	13457
Positive reciprocity	0.504	0.500	13457

Notes: Descriptive statistics for the full sample. Columns 1 and 2 report the mean and standard deviation, respectively. Column 3 reports the number of observations. Positive reciprocity, risk aversion, number of past jobs and income groups are based on the participants' national median levels.

B Additional Robustness

Table A.8: Intentions to freelance - excluding occupation 99

	(1)	(2)	(3)	(4)
Age (years)	-0.0638*** (0.0229)			-0.0608** (0.0253)
Age squared	0.0511* (0.0276)			0.0484 (0.0300)
Female	-0.3223*** (0.0732)			-0.2754*** (0.0734)
Educational attainment	-0.2181* (0.1324)			-0.1615 (0.1272)
Education Squared	0.0415 (0.0364)			0.0324 (0.0340)
Number of children	0.0244 (0.0305)			0.0135 (0.0308)
Migrant	0.0931 (0.1482)			0.0418 (0.1272)
Medium income	-0.1737* (0.0932)			-0.0709 (0.1032)
High income	-0.3453*** (0.0670)			-0.1508* (0.0790)
Risk averse		-0.2056*** (0.0522)		-0.1730*** (0.0424)
Internal locus		0.0450 (0.0661)		0.1572** (0.0786)
Low CRT		0.2310*** (0.0508)		0.1795*** (0.0580)
Positive reciprocity		0.2592*** (0.0844)		0.2241*** (0.0692)
Number of past jobs			-0.0003 (0.0075)	0.0172** (0.0081)
Manual-Creative			0.5174*** (0.1067)	0.4578*** (0.1004)
Knowledge-Routine			-0.1262** (0.0554)	-0.0586 (0.0574)
Knowledge-Creative			0.2029** (0.0892)	0.2350** (0.0923)
Atypical work arrangement			0.2382** (0.0971)	0.1890** (0.0906)
Private sector			0.2438*** (0.0649)	0.2293*** (0.0611)
Hours worked			-0.0330* (0.0196)	-0.0275 (0.0203)
Likely to lose job			1.3788*** (0.1220)	1.3788*** (0.1161)
Constant	0.3920 (0.4659)	-1.9978*** (0.0908)	-2.4033*** (0.2049)	-0.7628 (0.6106)
Observations	10744	10744	10744	10744
Pseudo R^2	0.0897	0.0814	0.1489	0.1633
BIC	8942	8976	8413	8273
Country F.E.	yes	yes	yes	yes
Occupation F.E.	no	no	yes	yes

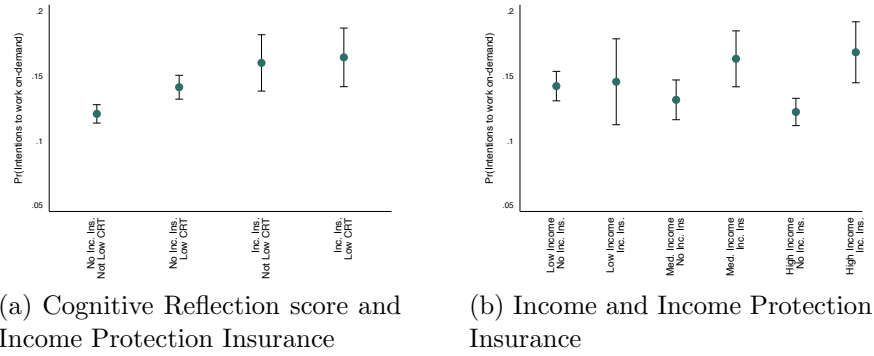
Notes: Logit regressions. Standard errors clustered at the country level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is a binary indicator taking value of one if the respondent reports being willing to engage in freelance work, and zero otherwise. Occupation dummies refer to the ISCO 2-digit classification of occupations. The reference category for task types is ‘Manual-Routine’ tasks.

Figure A.1: Intentions to Freelance by Age



Notes: Logistic odds of intentions to freelance (smoothed) by age. The regression model used includes all the same control variables as those presented in Table 1 column 4. The dependent variable is a binary indicator taking value of one if the respondent reports being willing to engage in freelance work, and zero otherwise. This graph reveals the non-linear effects of age discussed in the paper.

Figure A.2: Predicted probability of freelancing intentions by CRT score, Income bracket and possession income insurance



Notes: Predicted probability of freelancing for the four possible combinations of whether a person has a low CRT and possess income insurance (panel a) and 6 combinations of income and possession of income insurance (panel b). A low CRT is defined as a person failing all three CRT questions. Income insurance was defined as insurance that covers part of one's income if a person becomes unable to work due to illness or an accident. The dots represent marginal effects from the interaction variable, and the bars show 95% confidence intervals. All regressions include the full set of control as per Table 1 column (4).

Table A.9: Full model - Country-Task type Interaction Fixed Effects

	(1)
Age (years)	-0.1629*** (0.0162)
Age squared	0.1651*** (0.0187)
Female	0.0429 (0.0464)
Educational attainment	-0.1918** (0.0843)
Education squared	0.0589*** (0.0216)
Number of children	-0.0801*** (0.0257)
Migrant	-0.0997 (0.0671)
Medium income	-0.0646 (0.0529)
High income	-0.1295** (0.0641)
Risk averse	-0.1431*** (0.0416)
Internal Locus	-0.1321*** (0.0381)
Low CRT	-0.1125** (0.0455)
Positive reciprocity	0.0303 (0.0417)
Number of past jobs	0.0338*** (0.0050)
Manual-Creative	-0.2088*** (0.0461)
Knowledge-Routine	-0.2985*** (0.0259)
Knowledge-Creative	-0.2651*** (0.0351)
Atypical work arrangement	0.0236 (0.0600)
Private sector	0.2096*** (0.0690)
Hours worked	0.0028 (0.0157)
Likely to lose job	1.6083*** (0.1059)
Constant	2.7195*** (0.3810)
Observations	13454
Pseudo R^2	0.1306
BIC	13898
Task Type * Country Interaction F.E.	yes

Notes: Logit regression. Standard errors clustered at the country level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is a binary indicator taking value of one if the respondent reports being willing to engage in freelance work, and zero otherwise. Occupation dummies refer to the ISCO 2-digit classification of occupations. The reference category for task types is 'Manual-Routine' tasks.

Table A.10: Main Results: Only ‘Very Likely to Freelance’

	(1)	(2)	(3)	(4)
Age (years)	-0.1270*** (0.0206)			-0.1123*** (0.0210)
Age squared	0.1212*** (0.0247)			0.1028*** (0.0248)
Female	-0.3717*** (0.0696)			-0.3573*** (0.0737)
Educational attainment	-0.2614 (0.1932)			-0.2721 (0.1729)
Education Squared	0.0387 (0.0555)			0.0542 (0.0496)
Number of children	0.0511 (0.0444)			0.0475 (0.0428)
Migrant	-0.1953 (0.1472)			-0.1809* (0.0998)
Medium income	-0.3822** (0.1761)			-0.2604 (0.1738)
High income	-0.2710* (0.1469)			-0.1144 (0.1593)
Number of past jobs		0.0175 (0.0131)		0.0361*** (0.0117)
Manual-Creative		0.2112 (0.1603)		0.1318 (0.1663)
Knowledge-Routine		-0.1379 (0.1114)		-0.0573 (0.1263)
Knowledge-Creative		0.0315 (0.1137)		0.0811 (0.1336)
Atypical work arrangement		0.5365*** (0.1594)		0.4567*** (0.1436)
Private sector		0.2071* (0.1251)		0.1872 (0.1179)
Hours worked		-0.0389* (0.0220)		-0.0247 (0.0248)
Risk averse			-0.1443*** (0.0535)	-0.0819 (0.0589)
Internal LoC			0.1673 (0.1042)	0.1944** (0.0979)
Low CRT			0.1912 (0.1324)	0.1342 (0.1269)
Positive Reciprocity			0.3547*** (0.1208)	0.3023*** (0.1118)
Constant	0.3363 (0.4620)	-3.6869*** (0.1526)	-3.6376*** (0.1673)	-0.8949* (0.4938)
Observations	13457	13322	13457	13322
Pseudo R^2	0.0900	0.0933	0.0754	0.1147
BIC	4460	4490	4482	4387
Country F.E.	yes	yes	yes	yes
Occupation F.E.	no	yes	no	yes

Notes: Logit regressions. Standard errors clustered at the country level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is a binary indicator taking value of one if the respondent reports being very willing to engage in freelance work, and zero otherwise. Occupation dummies refer to the ISCO 2-digit classification of occupations. The reference category for task types is ‘Manual-Routine’ tasks.

Table A.11: Country-Level Risk Aversion Effects on Freelance Intentions

	Coefficient	Std. Error
Australia	-0.320	0.259
Brazil	-0.264	0.192
Germany	-1.183	0.404
Hong Kong	0.039	0.195
Ireland	-0.164	0.267
Italy	-0.218	0.231
Japan	-0.274	0.221
Malaysia	-0.252	0.192
Mexico	-0.025	0.191
Romania	-0.257	0.204
Spain	-0.523	0.330
Switzerland	-0.522	0.231
UAE	-0.203	0.155
UK	-0.382	0.282
US	-0.263	0.248
Finland	-0.259	0.304

Notes: The table displays the coefficients and standard errors for the binary indicator of Risk Aversion, from logit regressions run separately for each country. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is a binary indicator taking value of one if the respondent is estimated to have a risk aversion higher than their country median, and zero otherwise. All regressions include the full set of control variables as per Table 2 column (4).

Table A.12: Main Results with National Statistics

	(1)	(2)	(3)	(4)
Age (years)	-0.1546*** (0.0133)	-0.1600*** (0.0135)	-0.1561*** (0.0134)	-0.1601*** (0.0135)
Age squared	0.1552*** (0.0155)	0.1619*** (0.0157)	0.1580*** (0.0156)	0.1618*** (0.0157)
Female	0.0508 (0.0451)	0.0459 (0.0456)	0.0550 (0.0456)	0.0427 (0.0456)
Educational attainment	-0.1769** (0.0796)	-0.1979** (0.0808)	-0.1929** (0.0807)	-0.1967** (0.0806)
Education squared	0.0627*** (0.0232)	0.0605** (0.0235)	0.0608*** (0.0235)	0.0601** (0.0235)
Number of children	-0.0762*** (0.0211)	-0.0826*** (0.0215)	-0.0868*** (0.0214)	-0.0810*** (0.0215)
Migrant	0.0066 (0.0596)	-0.1009 (0.0690)	-0.0997 (0.0689)	-0.0785 (0.0634)
Medium income	-0.0249 (0.0544)	-0.0584 (0.0553)	-0.0403 (0.0553)	-0.0571 (0.0550)
High income	-0.1142** (0.0556)	-0.1226** (0.0564)	-0.1276** (0.0563)	-0.1221** (0.0564)
Risk averse	-0.1205*** (0.0430)	-0.1404*** (0.0435)	-0.1203*** (0.0433)	-0.1409*** (0.0434)
Internal Locus	-0.1100** (0.0433)	-0.1300*** (0.0438)	-0.1235*** (0.0438)	-0.1311*** (0.0438)
Low CRT	-0.1035** (0.0439)	-0.1126** (0.0446)	-0.1188*** (0.0446)	-0.1094** (0.0447)
Positive reciprocity	0.0263 (0.0428)	0.0329 (0.0431)	0.0333 (0.0430)	0.0326 (0.0431)
Number of past jobs	0.0262*** (0.0053)	0.0323*** (0.0054)	0.0293*** (0.0054)	0.0323*** (0.0054)
Manual-Creative	-0.1613** (0.0661)	-0.1671** (0.0666)	-0.1739*** (0.0665)	-0.1668** (0.0666)
Knowledge-Routine	-0.1463*** (0.0567)	-0.1364** (0.0573)	-0.1290** (0.0573)	-0.1373** (0.0571)
Knowledge-Creative	-0.0777 (0.0600)	-0.0669 (0.0607)	-0.0656 (0.0606)	-0.0680 (0.0606)
Atypical work arrangement	0.0560 (0.0485)	0.0274 (0.0509)	0.0248 (0.0513)	0.0292 (0.0508)
Private sector	0.1870*** (0.0449)	0.2098*** (0.0457)	0.2102*** (0.0456)	0.2112*** (0.0456)
Hours worked	-0.0018 (0.0163)	0.0032 (0.0166)	0.0051 (0.0165)	0.0027 (0.0166)
Likely to lose job	1.6213*** (0.0484)	1.6058*** (0.0493)	1.6011*** (0.0491)	1.6060*** (0.0492)
Constant	2.3870*** (0.3021)	2.6372*** (0.3124)	2.9196*** (0.3166)	2.8453*** (0.3171)
Observations	13454	13454	13454	13454
Pseudo R^2	0.1183	0.1273	0.1248	0.1273
BIC	14569	14521	14561	14511
Occupation F.E.	yes	yes	yes	yes
Global Rights F.E.	yes	no	no	no
Countr Priv. Share F.E.	no	yes	no	no
Countr Self-Employ Share F.E.	no	no	yes	no
Countr Self-Employ Growth F.E.	no	no	no	yes

Notes: Logit regressions. Standard errors clustered at the country level are reported in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. The dependent variable is a binary indicator taking value of one if the respondent reports being willing to engage in freelance work, and zero otherwise. Occupation dummies refer to the ISCO 2-digit classification of occupations. The reference category for task types is 'Manual-Routine' tasks. Global Rights dummy refers to ILO ITUC data.

C Appendix - Survey Questions

In this section, we provide the survey questions, as they were formulated (English translation) used in this study. We list the questions in order of appearance for the participants.

C.1 Screening Demographic Questions

- What is your gender? [Male; Female]
- What is your age (in years)?
- How would you describe your current employment?
[I am employed (i.e., work for an organization or company that is not my own); I am self-employed (e.g., have my own business, am a freelancer, work-on-demand); Both of the above]
- What is your income (gross)? [the typical reference period (week, month, year) and income brackets were chosen by the survey provider, specifically for each country.]

C.2 Work Situation Questions

- How many hours do you work per week? [Less than 10 hours; 10 to 14 hours; 15 to 20 hours; 21 to 34 hours; 35 to 40 hours; More than 40 hours]
- How would you describe your (main) job? (open text)
- What type of organization are you working for? [A privately owned company; A publicly traded company; A government organisation; A cooperative; An NGO/charity/non-profit organisation; I don't know]

- In your (main) job do you have the ability to influence how your daily work is organised? [I have no influence; I have some influence; I have influence; I have a lot of influence; I have complete control]
- In your (main) job do you have the ability influence important decisions regarding the organization's activities(e.g.,involving its strategy)? [I have no influence; I have some influence; I have influence; I have a lot of influence; I have complete control]
- Thinking about your (main) job, which percent of your time do you spent on manual and on knowledge-based tasks? [2 Likert scales from 0 to 100 were offered for each of manual and knowledge-based tasks].
- Thinking about your (main) job, which percent of your time do you spent on routine and on creative tasks? [2 Likert scales from 0 to 100 were offered for each of routine and creative tasks].
- Which type of employment contract do you have for your (main) job? [I have a contract for a fixed amount of time: for l e s s than 3 months; I have a contract for a fixed amount of time: for m o r e than 3 months; I have a permanent contract (i.e.,of unlimited duration); I have a contract for an apprenticeship (or other training contract); I do not have any contract]
- How many jobs have you had since the beginning of your career? (Please count jobs held for more than 3 months and include your current, part-time and zero hour contract jobs. If you are or have been self-employed count it as one job].)

C.3 Career Questions

- Thinking about the next 12 months, how likely do you think you are to lose your (main) job? [Very unlikely; Somewhat unlikely; Somewhat likely;

Very likely]

- Thinking about the next 12 months, how likely do you think you are to leave your (main) job voluntarily [4-point: Very unlikely to Very likely]
- Thinking about the next 12 months, how likely do you think you are to leave your (main) job to become a freelancer (become self-employed and work on demand)? [Very unlikely; Somewhat unlikely; Somewhat likely; Very likely]

C.4 Individual Attitudes Questions

- A bat and a ball cost \$11.00 in total. The bat costs \$10.00 more than the ball. How much does the ball cost? (Number Input: 0-200)
- If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? (Number Input: 0-200)
- In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? (Number Input: 0-200)
- Some people feel they have a great deal of control over their lives, while other people feel that what they do has no real effect on what happens to them. How much control do you feel you have? [5-point-scale: No control to Great deal of control]
- Some people feel they have complete freedom of choice in their lives, while other people do not. How much freedom of choice do you feel you have? [5-point-scale from no control to a great deal of control]
- Imagine the following situation: Today you unexpectedly received \$1,000. You can choose to keep it yourself or donate it (in part to a good cause).

How much of this amount would you donate to a good cause? [slider range: \$0 – \$1000]

- Imagine that you have a \$100 of savings. You have the opportunity to invest in a risky financial product with the following features: There is a chance of 1 in 2 (50%) that the investment is profitable. If the investment is profitable, you earn 2.5 times the amount that you have invested. If the investment is not profitable, you lose the amount that you have invested.
- How much of your saved \$100 would you like to invest in this product? [Slider range: \$0 – \$100]

C.5 Final Demographic Questions

- How many years of schooling have you completed in your life? [Range: 1-25 years]
- Do you have any children? [Yes; No]
- How many children do you have? [Range: 1-15]
- Were you born in your current country of residence? [Yes; No]
- All in all, how would you describe your current health state if you compare yourself to other people of your own age? [A lot worse than average; Worse than average; Average; Better than average; A lot better than average]

References

- Acemoglu, D. and Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In *Handbook of Labor Economics*, volume 4, pages 1043–1171. Elsevier.
- Acs, Z. J., Audretsch, D. B., and Evans, D. S. (1994). Why does the self-employment rate vary across countries and over time? Technical report, CEPR Discussion Papers.
- Audretsch, D. B., Kuratko, D. F., and Link, A. N. (2015). Making sense of the elusive paradigm of entrepreneurship. *Small business economics*, 45:703–712.
- Baron, R. A. and Markman, G. D. (2000). Beyond social capital: How social skills can enhance entrepreneurs’ success. *Academy of Management Perspectives*, 14(1):106–116.
- Bates, T. (1995). Self-employment entry across industry groups. *Journal of Business Venturing*, 10(2):143–156.
- Beugelsdijk, S. and Noorderhaven, N. (2005). Personality characteristics of self-employed; an empirical study. *Small Business Economics*, 24:159–167.
- Bird, B. (1988). Implementing entrepreneurial ideas: The case for intention. *Academy of Management Review*, 13(3):442–453.
- Blanchflower, D. G. and Meyer, B. D. (1994). A longitudinal analysis of the young self-employed in australia and the united states. *Small Business Economics*, 6:1–19.
- Blanchflower, D. G. and Oswald, A. J. (1998). What makes an entrepreneur? *Journal of Labor Economics*, 16(1):26–60.
- Block, J. and Sandner, P. (2009). Necessity and opportunity entrepreneurs and their duration in self-employment: evidence from german micro data. *Journal of Industry, Competition and Trade*, 9:117–137.
- Borjas, G. J. (1986). The self-employment experience of immigrants. *The Journal of Human Resources*, 21(4):485–506.
- Brockhaus Sr, R. H. (1980). Risk taking propensity of entrepreneurs. *Academy of Management Journal*, 23(3):509–520.
- Burke, A. and Cowling, M. (2020a). On the critical role of freelancers in agile economies. *Small Business Economics*, 55(2):393–398.
- Burke, A. and Cowling, M. (2020b). The role of freelancers in entrepreneurship and small business. *Small Business Economics*, 55(2):389–392.
- Cahill, K. E., Giandrea, M. D., and Quinn, J. F. (2013). New evidence on self-employment transitions among older americans with career jobs. Technical report, Bureau of Labor Statistics.

- Caliendo, M., Fossen, F., and Kritikos, A. (2012). Trust, positive reciprocity, and negative reciprocity: Do these traits impact entrepreneurial dynamics? *Journal of Economic Psychology*, 33(2):394–409.
- Caliendo, M., Fossen, F., and Kritikos, A. S. (2014). Personality characteristics and the decisions to become and stay self-employed. *Small Business Economics*, 42:787–814.
- Caliendo, M., Fossen, F. M., and Kritikos, A. S. (2009). Risk attitudes of nascent entrepreneurs—new evidence from an experimentally validated survey. *Small Business Economics*, 32:153–167.
- Chandler, G. N. and Hanks, S. H. (1998). An examination of the substitutability of founders human and financial capital in emerging business ventures. *Journal of Business Venturing*, 13(5):353–369.
- Charness, G. and Gneezy, U. (2010). Portfolio choice and risk attitudes: An experiment. *Economic Inquiry*, 48(1):133–146.
- Charness, G. and Villeval, M.-C. (2009). Cooperation and competition in inter-generational experiments in the field and the laboratory. *American Economic Review*, 99(3):956–978.
- Clain, S. H. (2000). Gender differences in full-time self-employment. *Journal of Economics and Business*, 52(6):499–513.
- Clark, A. E. (1997). Job satisfaction and gender: why are women so happy at work? *Labour Economics*, 4(4):341–372.
- Clark, G. L., Fiaschetti, M., and Gerrans, P. (2019). Determinants of seeking advice within defined contribution retirement savings schemes. *Accounting & Finance*, 59:563–591.
- Clark, K. and Drinkwater, S. (2000). Pushed out or pulled in? self-employment among ethnic minorities in england and wales. *Labour Economics*, 7(5):603–628.
- Croson, R. and Gneezy, U. (2009). Gender differences in preferences. *Journal of Economic Literature*, 47(2):448–474.
- Datta, N., Giupponi, G., and Machin, S. (2019). Zero-hours contracts and labour market policy. *Economic Policy*, 34(99):369–427.
- Dohmen, T., Falk, A., Huffman, D., and Sunde, U. (2010). Are risk aversion and impatience related to cognitive ability? *American Economic Review*, 100(3):1238–1260.
- Ekelund, J., Johansson, E., Järvelin, M.-R., and Lichtermann, D. (2005). Self-employment and risk aversion—evidence from psychological test data. *Labour Economics*, 12(5):649–659.

- Eurofound (2017). Atypical work. Technical report, European Industrial Relations Dictionary.
- Evans, D. S. and Leighton, L. S. (1989). The determinants of changes in us self-employment, 1968–1987. *Small Business Economics*, 1:111–119.
- Evans, D. S. and Leighton, L. S. (1990). Some empirical aspects of entrepreneurship. *The Economics of Small Firms*, pages 79–99.
- Fairlie, R. W. and Krashinsky, H. A. (2012). Liquidity constraints, household wealth, and entrepreneurship revisited. *Review of Income and Wealth*, 58(2):279–306.
- Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., and Sunde, U. (2018). Global evidence on economic preferences. *The Quarterly Journal of Economics*, 133(4):1645–1692.
- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4):25–42.
- Georgellis, Y., Sessions, J. G., and Tsitsianis, N. N. (2005). Self-employment longitudinal dynamics: A review of the literature. *Economic Issues*, 10(2).
- Georgellis, Y. and Wall, H. J. (2005). Gender differences in self-employment. *International Review of Applied Economics*, 19(3):321–342.
- Graham, M., Lehdonvirta, V., Wood, A., Barnard, H., Hjorth, I., and D Simon, P. (2017). The risks and rewards of online gig work at the global margins. *Oxford Internet Institute*.
- Grubb, D. and Wells, W. (1993). Employment regulation and patterns of work in ec countries. *OECD Economic Studies*, pages 7–7.
- Guerra, G. and Patuelli, R. (2016). The role of job satisfaction in transitions into self-employment. *Entrepreneurship Theory and practice*, 40(3):543–571.
- Hamilton, B. H. (2000). Does entrepreneurship pay? an empirical analysis of the returns to self-employment. *Journal of Political Economy*, 108(3):604–631.
- Handy, C. B. (1984). *The future of work: A guide to a changing society*. John Wiley & Sons.
- Hansemark, O. C. (2003). Need for achievement, locus of control and the prediction of business start-ups: A longitudinal study. *Journal of Economic Psychology*, 24(3):301–319.
- Henrekson, M. and Stenkula, M. (2010). Entrepreneurship and public policy. In *Handbook of entrepreneurship research: An interdisciplinary survey and introduction*, pages 595–637. Springer.

- Hintermaier, T. and Steinberger, T. (2005). Occupational choice and the private equity premium puzzle. *Journal of Economic Dynamics and Control*, 29(10):1765–1783.
- Hundley, G. (2006). Family background and the propensity for self-employment. *Industrial Relations: A Journal of Economy and Society*, 45(3):377–392.
- Hyytinen, A., Ilmakunnas, P., and Toivanen, O. (2013). The return-to- entrepreneurship puzzle. *Labour Economics*, 20:57–67.
- Inglehart, R., Foa, R., Peterson, C., and Welzel, C. (2008). Development, freedom, and rising happiness: A global perspective (1981–2007). *Perspectives on Psychological Science*, 3(4):264–285.
- Innocenti, S., Clark, G. L., and McGill, S. (2023). Experience of financial challenges, retirement concerns, and planning: evidence from representative samples of workers in 16 countries. *Journal of Pension Economics & Finance*, pages 1–19.
- Innocenti, S. and Golin, M. (2022). Human capital investment and perceived automation risks: Evidence from 16 countries. *Journal of Economic Behavior & Organization*, 195:27–41.
- Jones, M. K. and Latreille, P. L. (2011). Disability and self-employment: evidence for the uk. *Applied Economics*, 43(27):4161–4178.
- Judge, T. A. and Bono, J. E. (2001). Relationship of core self-evaluations traits—self-esteem, generalized self-efficacy, locus of control, and emotional stability—with job satisfaction and job performance: A meta-analysis. *Journal of Applied Psychology*, 86(1):80.
- Kässi, O. and Lehdonvirta, V. (2018). Online labour index: Measuring the online gig economy for policy and research. *Technological Forecasting and Social Change*, 137:241–248.
- Kihlstrom, R. E. and Laffont, J.-J. (1979). A general equilibrium entrepreneurial theory of firm formation based on risk aversion. *Journal of Political Economy*, 87(4):719–748.
- Kim, P. H., Aldrich, H. E., and Keister, L. A. (2006). Access (not) denied: The impact of financial, human, and cultural capital on entrepreneurial entry in the united states. *Small Business Economics*, 27:5–22.
- Kitching, J. and Smallbone, D. (2012). Are freelancers a neglected form of small business? *Journal of Small Business and Enterprise Development*, 19(1):74–91.
- Koellinger, P., Minniti, M., and Schade, C. (2013). Gender differences in entrepreneurial propensity. *Oxford Bulletin of Economics and Statistics*, 75(2):213–234.

- Kolvereid, L. and Isaksen, E. (2006). New business start-up and subsequent entry into self-employment. *Journal of Business Venturing*, 21(6):866–885.
- Krueger Jr, N. F., Reilly, M. D., and Carsrud, A. L. (2000). Competing models of entrepreneurial intentions. *Journal of Business Venturing*, 15(5-6):411–432.
- Lawrence, A., Clark, L., Labuzetta, J. N., Sahakian, B., and Vyakarnum, S. (2008). The innovative brain. *Nature*, 456(7219):168–169.
- Lin, Z., Picot, G., and Compton, J. (2000). The entry and exit dynamics of self-employment in canada. *Small Business Economics*, 15:105–125.
- Lofstrom, M., Bates, T., and Parker, S. C. (2014). Why are some people more likely to become small-businesses owners than others: Entrepreneurship entry and industry-specific barriers. *Journal of Business Venturing*, 29(2):232–251.
- MacDonald, R. and Giazitzoglu, A. (2019). Youth, enterprise and precarity: or, what is, and what is wrong with, the ‘gig economy’? *Journal of Sociology*, 55(4):724–740.
- Mueller, S. L. and Thomas, A. S. (2001). Culture and entrepreneurial potential: A nine country study of locus of control and innovativeness. *Journal of Business Venturing*, 16(1):51–75.
- Ng, T. W., Sorensen, K. L., and Eby, L. T. (2006). Locus of control at work: a meta-analysis. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, 27(8):1057–1087.
- Nieß, C. and Zacher, H. (2015). Openness to experience as a predictor and outcome of upward job changes into managerial and professional positions. *PloS one*, 10(6):e0131115.
- Nikolaev, B., Boudreaux, C. J., and Wood, M. (2020). Entrepreneurship and subjective well-being: The mediating role of psychological functioning. *Entrepreneurship Theory and Practice*, 44(3):557–586.
- Nykqvist, J. (2008). Entrepreneurship and liquidity constraints: Evidence from sweden. *Scandinavian Journal of Economics*, 110(1):23–43.
- Obschonka, M., Silbereisen, R. K., and Schmitt-Rodermund, E. (2010). Entrepreneurial intention as developmental outcome. *Journal of Vocational Behavior*, 77(1):63–72.
- Pagán, R. (2009). Self-employment among people with disabilities: evidence for europe. *Disability & Society*, 24(2):217–229.
- Parker, S. C. (2004). *The economics of self-employment and entrepreneurship*. Cambridge University Press.

- Parker, S. C. and Rougier, J. C. (2007). The retirement behaviour of the self-employed in Britain. *Applied Economics*, 39(6):697–713.
- Poschke, M. (2013). Who becomes an entrepreneur? labor market prospects and occupational choice. *Journal of Economic Dynamics and Control*, 37(3):693–710.
- Prassl, J. (2018). *Humans as a service: The promise and perils of work in the gig economy*. Oxford University Press.
- Prassl, J. and Risak, M. (2015). Uber, taskrabbit, and co.: Platforms as employers-rethinking the legal analysis of crowdwork. *Comp. Lab. L. & Pol’y J.*, 37:619.
- Rees, H. and Shah, A. (1986). An empirical analysis of self-employment in the UK. *Journal of Applied Econometrics*, 1(1):95–108.
- Scarelli, T. and Margolis, D. N. (2023). When you can’t afford to wait for a job: The role of time discounting for own-account workers in developing countries. *IZA Discussion Paper No. 15926*.
- Schmitt-Rodermund, E. (2004). Pathways to successful entrepreneurship: Parenting, personality, early entrepreneurial competence, and interests. *Journal of Vocational Behavior*, 65(3):498–518.
- Simoes, N., Crespo, N., and Moreira, S. B. (2016). Individual determinants of self-employment entry: What do we really know? *Journal of Economic Surveys*, 30(4):783–806.
- Stefanović, S. and Stošić, D. (2012). Specifics and challenges of female entrepreneurship. *Economic Themes*, 50(3).
- Stephan, U., Tavares, S. M., Carvalho, H., Ramalho, J. J., Santos, S. C., and Van Veldhoven, M. (2020). Self-employment and eudaimonic well-being: Energized by meaning, enabled by societal legitimacy. *Journal of Business Venturing*, 35(6):106047.
- Stewart Jr, W. H. and Roth, P. L. (2001). Risk propensity differences between entrepreneurs and managers: A meta-analytic review. *Journal of Applied Psychology*, 86(1):145.
- Stewart Jr, W. H., Watson, W. E., Carland, J. C., and Carland, J. W. (1999). A proclivity for entrepreneurship: A comparison of entrepreneurs, small business owners, and corporate managers. *Journal of Business Venturing*, 14(2):189–214.
- Taylor, M. P. (2001). Self-employment and windfall gains in Britain: evidence from panel data. *Economica*, 68(272):539–565.
- Torrini, R. (2005). Cross-country differences in self-employment rates: The role of institutions. *Labour Economics*, 12(5):661–683.

- Van den Born, A. and Van Witteloostuijn, A. (2013). Drivers of freelance career success. *Journal of Organizational Behavior*, 34(1):24–46.
- Van der Sluis, J., Van Praag, M., and Vijverberg, W. (2008). Education and entrepreneurship selection and performance: A review of the empirical literature. *Journal of Economic Surveys*, 22(5):795–841.
- van Gelderen, M., Wiklund, J., and McMullen, J. S. (2021). Entrepreneurship in the future: A delphi study of etp and jbv editorial board members. *Entrepreneurship Theory and Practice*, 45(5):1239–1275.
- Veenhoven, R. (2000). Freedom and happiness: A comparative study in 46 nations in the early 1990s. In: *Diener, E. & Suh, E.M. (eds) 'Culture and Subjective Wellbeing'*, pages 257–288.
- Vejsiu, A. (2011). Incentives to self-employment decision in sweden. *International Review of Applied Economics*, 25(4):379–403.
- Verme, P. (2009). Happiness, freedom and control. *Journal of Economic Behavior & Organization*, 71(2):146–161.
- Wagner, J. (2007). What a difference a y makes-female and male nascent entrepreneurs in germany. *Small Business Economics*, 28:1–21.
- Williams, D. R. (2012). Gender discrimination and self-employment dynamics in europe. *The Journal of Socio-Economics*, 41(2):153–158.
- Williamson, C. R. and Mathers, R. L. (2011). Economic freedom, culture, and growth. *Public Choice*, 148:313–335.
- Wood, A. J. (2016). Flexible scheduling, degradation of job quality and barriers to collective voice. *Human Relations*, 69(10):1989–2010.
- Wood, A. J., Graham, M., Lehdonvirta, V., and Hjorth, I. (2019). Good gig, bad gig: autonomy and algorithmic control in the global gig economy. *Work, Employment and Society*, 33(1):56–75.
- Woodcock, J. and Graham, M. (2019). *The gig economy*. A critical introduction. Cambridge: Polity.
- Woronkiewicz, J. and Noonan, D. S. (2019). Who goes freelance? the determinants of self-employment for artists. *Entrepreneurship Theory and Practice*, 43(4):651–672.
- Wyrwich, M. (2013). Can socioeconomic heritage produce a lost generation with regard to entrepreneurship? *Journal of Business Venturing*, 28(5):667–682.
- Zissimopoulos, J. M. and Karoly, L. A. (2007). Transitions to self-employment at older ages: The role of wealth, health, health insurance and other factors. *Labour Economics*, 14(2):269–295.

Chapter 4 | Paper 2

Do Food Subsidies Change Diets for Good? Evidence from Subsidy Introduction and Withdrawal

Abstract: Price subsidies are widely proposed to promote lower-carbon diets, yet real-world evidence on their effects remains scarce, and whether any effect persists once the subsidy is withdrawn is largely untested. Using two natural experiments in university cafeterias, we study the introduction and withdrawal of price subsidies on vegetarian meals, distinguishing temporary price responses from durable behavioural change in a high-frequency, repeated choice setting. One cafeteria introduced a 5% subsidy on vegetarian meals, while another removed a pre-existing 14% subsidy. Analysing over 110,000 meal choices, we find that introducing the subsidy raises vegetarian uptake by 3 percentage points, while removing one reduces it by a comparable amount. In both experiments, responses are concentrated among consumers with more vegetarian baseline diets. Event-study estimates show that behaviour adjusts immediately at each policy change, and the withdrawal experiment provides no evidence of persistence once the subsidy is removed. These findings are more consistent with heterogeneous price responsiveness than with habit formation. Back-of-the-envelope calculations imply an abatement cost of roughly £1,180 per tonne of CO₂e, an order of magnitude above standard social cost of carbon estimates and well above comparable non-price interventions such as carbon labelling. Food subsidies thus lower the carbon content of diets while incentives are in place, but do not appear to produce lasting behavioural change.

Declaration of Authorship: I declare, as per University of Oxford School of Geography and the Environment regulations, that I have contributed more than the required 50% to the intellectual content of this paper, and that I am the

primary author.

Authors: Adam Ferris, Paul Lohmann, and Andrea Byfuglien.

The paper is included in the next pages.

Do Food Subsidies Change Diets for Good? Evidence from Subsidy Introduction and Withdrawal

Adam Ferris, Paul Lohmann, and Andrea Byfuglien

Abstract

Price subsidies are widely proposed to promote lower-carbon diets, yet real-world evidence on their effects remains scarce, and whether any effect persists once the subsidy is withdrawn is largely untested. Using two natural experiments in university cafeterias, we study the introduction and withdrawal of price subsidies on vegetarian meals, distinguishing temporary price responses from durable behavioural change in a high-frequency, repeated choice setting. One cafeteria introduced a 5% subsidy on vegetarian meals, while another removed a pre-existing 14% subsidy. Analysing over 110,000 meal choices, we find that introducing the subsidy raises vegetarian uptake by 3 percentage points, while removing one reduces it by a comparable amount. In both experiments, responses are concentrated among consumers with more vegetarian baseline diets. Event-study estimates show that behaviour adjusts immediately at each policy change, and the withdrawal experiment provides no evidence of persistence once the subsidy is removed. These findings are more consistent with heterogeneous price responsiveness than with habit formation. Back-of-the-envelope calculations imply an abatement cost of roughly £1,180 per tonne of CO₂e, an order of magnitude above standard social cost of carbon estimates and well above comparable non-price interventions such as carbon labelling. Food subsidies thus lower the carbon content of diets while incentives are in place, but do not appear to produce lasting behavioural change.

Keywords: Price Subsidies, Policy Persistence, Habit Formation, Price Responsiveness, Sustainable Consumption

JEL Codes: C23, D12, H23, Q18, Q58

1 Introduction

Pricing instruments, such as taxes and subsidies, are the workhorses of climate policy. They leverage market forces through price-based incentives, and have been widely adopted to shape household energy use, travel behaviour, and food choices at scale. At the same time, they remain a point of debate in the climate policy sphere (Colmer et al., 2025; Dechezleprêtre et al., 2023; Clinton and Steinberg, 2019; Acemoglu et al., 2012; van der Ploeg, 2025; Fageda and Teixidó, 2022; Douenne and Fabre, 2022). A central question in this debate is whether temporary price incentives produce lasting changes in behaviour, or merely encourage short-run behavioural responses that disappear once the incentive is withdrawn (Gneezy et al., 2011; Charness and Gneezy, 2009; Royer et al., 2015; Carrera et al., 2020). The distinction matters considerably for cost-effectiveness and policy valuation. If a subsidy needs to be maintained indefinitely to sustain its effect, the cumulative incentive cost may be larger than the social benefit of the incentivised behaviour; if instead the subsidy reshapes underlying preferences or habits, even a temporary programme can deliver durable returns (Becker and Murphy, 1988; Bowles and Polania-Reyes, 2012; Mattauch et al., 2022).

About two-thirds of global emissions are linked to consumer-driven sources of emissions (Ivanova et al., 2020; Clark et al., 2020). One of the largest of these sources – what people eat – remains comparatively under-examined in the context of price policy, despite persistent calls for demand-side mitigation (Creutzig et al., 2022; Lohmann et al., 2026).¹ Unlike many consumer decisions, dietary choices are shaped by cultural identity, habit, and social norms, which can make sustained behavioural change difficult to achieve. At the same time, recent structural demand

¹Though the literature is considerably larger for food policy and health; see Griffith (2022) for a review. Much of the broader demand-side literature on sustainable food consumption has focused on non-price levers such as nudges, defaults, and information provision (see e.g. Kurz, 2018; Dannenberg and Weingärtner, 2023; Dannenberg et al., 2024; Fosgaard et al., 2024; Panzone et al., 2024; Lohmann et al., 2022; Perino and Schwirplies, 2022).

estimates suggest that pricing policies, primarily carbon taxes, can meaningfully shift food consumption and welfare (Kehlbacher et al., 2016; Tomasi et al., 2026), offering a useful benchmark for the scale of price responsiveness in this domain. Whether comparable subsidies for lower-carbon alternatives generate similar effects, and whether any such response persists after prices return to their original level, remains an open question.

Theory offers competing predictions for how consumers respond when price incentives are introduced and when they are withdrawn. Standard price theory predicts heterogeneous responses concentrated among marginal consumers, with limited persistence after removal. By contrast, endogenous preference and habit-formation models imply that sustained incentive-induced behaviour may persist even after incentives are withdrawn (Bowles and Polania-Reyes, 2012; Mattauch et al., 2022; Holt and Goodwin, 1997; Carrasco et al., 2005). Reference dependence and salience models additionally imply that introduction and withdrawal need not be symmetric (Kőszegi and Rabin, 2006; DellaVigna et al., 2017; Bordalo et al., 2013, 2020). Crucially, distinguishing these theories requires observing what happens after policy withdrawal: persistence is consistent with habit formation, while reversion is consistent with ongoing price responsiveness. Designs that observe only the introduction of an incentive cannot separate these channels.

This paper studies both stages of this policy lifecycle through two natural experiments in university cafeterias: in one cafeteria, a 5% subsidy on vegetarian meals is introduced (the *subsidy-introduction experiment*); in another, a pre-existing 14% subsidy is removed (the *subsidy-withdrawal experiment*). Two additional cafeterias with stable pricing provide control trajectories. Policymakers routinely face both decisions: launching a programme under budget constraints and, later, deciding whether it has earned its keep (Gneezy et al., 2011; Thaler, 2018). The cafeteria setting is well-suited to distinguishing price responsiveness from habit formation,

as individuals face the same repeated choice several times per week under stable institutional conditions - precisely the circumstances that endogenous preference and habit-formation models require in order to generate lasting change. Observing both introduction and withdrawal is what allows us to separate these mechanisms: if habits formed during the subsidy period, removal should have little effect; if behaviour was price-contingent throughout, it should reverse. We estimate difference-in-differences models with individual, college-by-mealtime, and week fixed effects, and study heterogeneity by pre-intervention consumption shares. The subsidy-introduction experiment was pre-registered. The subsidy-withdrawal experiment was identified during data curation and analysed using the same pre-registered empirical strategy.²

Three findings emerge from this research. First, introducing a 5% subsidy on vegetarian meals raises vegetarian meal uptake by roughly 3 percentage points (pp) on average, with effects concentrated among students who already chose vegetarian meals predominantly or somewhat frequently before the policy (around 6 pp for these groups). Second, withdrawing a 14% subsidy yields a mirroring 3 pp decrease in vegetarian meal demand, again driven by students who more frequently chose vegetarian meals at baseline (a 6.7 pp reduction among flexitarian consumers), while those who rarely chose them are unaffected.³ Event-study evidence indicates that responses are immediate rather than gradual, with no sign of self-sustaining behavioural shifts once the policy is withdrawn. Third, back-of-the-envelope calculations show that the 5% subsidy generates only a modest emissions reduction of around 2.6% relative to baseline, implying limited cost-effectiveness

²At pre-registration we anticipated two cafeterias applying vegetarian subsidies over the study period. On examining the actual meal prices in the sales data, we found that one had instead removed a pre-existing subsidy at the start of the second year, giving rise to the subsidy-withdrawal experiment. Because this variation could not have been anticipated before accessing the data, we report the withdrawal results as exploratory and apply to them the specification pre-registered for the subsidy-introduction experiment. We document all departures from the pre-analysis plan in Appendix A.1.

³The baseline year for the subsidy-withdrawal experiment is the year the subsidy still existed.

with an abatement cost of approximately £1,180 per tonne of CO₂e, substantially exceeding benchmark estimates of the social cost of carbon and reported abatement costs for non-price interventions such as cafeteria carbon-labelling schemes (Moore et al., 2024; Lohmann et al., 2022; Schulze-Tilling, 2025).

We contribute to three strands of literature. First, we add to the evidence on the effectiveness of taxes and subsidies in shaping behaviour across policy domains (Parry and Small, 2005; Metcalf, 2021; Bowles and Polania-Reyes, 2012; Chetty et al., 2009), and to the growing literature on the behavioural economics of price-based climate policy (Gsottbauer and van den Bergh, 2025; Andor et al., 2023; Bartels and Werthschulte, 2025). Second, we contribute to the literature on food pricing, bringing comparable real-world, longitudinal evidence to a setting where price-based instruments are seldom evaluated. Prior work, largely short-horizon or correlational, shows mixed effects of price interventions on healthier or lower-carbon choices (An, 2013; Mytton et al., 2012; Garnett et al., 2021; Panzone et al., 2021b,a; Pizzo et al., 2024; Lohmann et al., 2024) and highlights acceptability constraints relative to taxes (Pechey et al., 2022c; Lehmann et al., 2025; Funke et al., 2022). We provide field evidence that offers high internal validity by limiting cross-venue substitution, while observing revealed food choices in a price-sensitive student population. Third, evidence on what happens after price-based interventions are withdrawn remains limited, and concentrated outside the food domain – for example in payments for ecosystem services (Kemigisha et al., 2023; Jayachandran et al., 2018) and free or reduced public transport (Gravert and Collentine, 2021; Bull et al., 2021). By observing both subsidy introduction and withdrawal in comparable institutional settings, we provide a direct field test of whether incentive-induced behaviour persists after policy removal, a test that most single-policy studies cannot deliver (Royer et al., 2015; Carrera et al., 2020).

The rest of the paper is organised as follows. Section 2 describes the setting, data,

and estimation strategy. Section 3 presents main treatment effects for subsidy introduction and withdrawal. Section 4 analyses heterogeneity by baseline diet composition. Section 5 studies dynamic responses through event-study evidence. Section 6 concludes.

2 Methods and Data

2.1 Experimental Design

We study the effect of a cafeteria-level subsidy on vegetarian meals through two natural experiments conducted across four cafeterias at the University of Oxford.⁴ The experiment spans two academic years, from September 2021 to July 2023, with data collected from anonymised individual-level meal sales. College cafeterias are comparable to residential dining halls, offering daily lunches and dinner services to its members, including students, faculty, staff and occasional guests. However, our analysis focuses exclusively on student meal choices.⁵

The four colleges host between 685 and 770 students annually, admitting both undergraduate and graduate students (Master’s and PhD-level degrees) across all academic disciplines. Each cafeteria provides lunch services to all student residents on all weekdays, as well as dinner services on most weekdays. While the exact menu composition and number of dishes offered varied daily, all cafeterias consistently provide vegan/vegetarian, meat, and fish options during lunch and dinner. Menus are independently determined by each cafeteria, with chefs ensuring the continuous availability of these three categories of dishes throughout meal services. Our study

⁴Ethical approval for this study was granted by the University of Oxford, CUREC: SOGE C1A 23 103.

The subsidy-introduction analysis was pre-registered before we accessed and curated the full dataset (https://osf.io/xqg5f/overview?view_only=12af6b486b41406990057eae08574357, October 2024). The subsidy-withdrawal analysis was not pre-registered and should be read as exploratory; we document all departures from the pre-analysis plan in Appendix A.1.

⁵Staff and faculty are excluded for our analysis because they are subject to different pricing schemes, and have access to different meal options in some cases.

builds on prior research that leverages college cafeteria sales data, which provides high-quality, granular information on individual meal-purchasing behaviour over time, enabling robust empirical analysis (Garnett et al., 2019, 2021; Lohmann et al., 2022).

We obtained anonymised individual-level meal sales data from all cafeterias, covering two academic years from September 2021 to July 2023. Each academic year consists of three 8-week terms, during which most students reside in college accommodation.⁶ Sales data were recorded electronically, with each transaction linked to a consistent unique ID number when students paid using their University ID card. This payment method is the most common among students, as it provides access to subsidised meal rates across all dishes for college members. Our analysis focuses exclusively on students who paid using their student ID cards.⁷ Students are unable to use their student cards or access subsidised meals at cafeterias outside of their own college, and they rarely invite students from other colleges to dine with them.⁸ The unique IDs allow us to track each student's meal choices over at least two academic years, capturing detailed information on the date, time and whether the meal was vegetarian).

Most students have access to fully equipped kitchens and external food outlets (e.g. restaurants, pubs, takeaways, and supermarkets). However, these alternatives are typically much more expensive than the cafeteria meals, which are heavily subsidised by the colleges. Lunches at the cafeterias cost approximately £2–4, while dinners are priced around £3–4.⁹ College cafeterias also serve as a social setting

⁶The academic terms broadly align with seasons: the autumn term runs from October to December, the winter term from January to March, and the spring term from April to June. Our analysis is restricted to term-time data with an additional week at the beginning of term, as undergraduates typically arrive a few days before the official start of term and are required to vacate college accommodation in the last week of term time, resulting in changes to the sample composition.

⁷Each year, new students join the colleges while others graduate and leave, leading to changes in sample composition. However, our primary estimation strategy, which employs student-level fixed effects, uses data only from students observed both before and after the intervention period.

⁸We exclude sales where a university ID card is associated with two or more purchases at the same mealtime, as these may reflect purchases made for guests.

⁹A table of all college meal prices from 2021–2023 is provided in Table A.12

where students gather with peers from their colleges, further contributing to their popularity.

2.2 Intervention

We leverage a natural experiment involving two pricing policy occurrences at two of the cafeterias: (1) The introduction of a vegetarian meal subsidy policy at one cafeteria, and (2) the withdrawal of a vegetarian meal subsidy policy at another. The academic year spanning October 2021 to June 2022 serves as the baseline across all four cafeterias, with data from one subsequent academic year (2022–23) included in the analysis. Since the policy changes (subsidy introduction and withdrawal) were implemented at the start of the 2022–23 academic year, our primary analysis focuses on comparing sales data from 2021–22 (pre-policy baseline) and 2022–23 (post-policy), creating a clean pre-post identification.

During the baseline year (2021–22), College A did not differentiate prices between vegetarian and non-vegetarian meals, while Colleges B, C and D offered subsidies on vegetarian options. Specifically, College B, C and D provided a 14%, 5% and a 28% subsidy, respectively, on vegetarian meals relative to meat and fish options (see Table 1).¹⁰ Colleges C and D maintained their subsidies throughout the study period (pre- and post-policy), serving as the control cafeterias for unmeasured time trends and cycles.

College B removed its vegetarian meal subsidy at the start of the second academic year (2022–23), making it the treatment cafeteria for the subsidy withdrawal study. College B also maintained a price premium on fish dishes across both years. Meanwhile, College A introduced a 5% subsidy on vegetarian meals in 2022–23, thus serving as the treatment cafeteria for the subsidy introduction analysis. A summary of lunch price changes is presented in Table 1, with further details provided

¹⁰College B only provided this subsidy type for lunch meals.

in the appendix (Table A.12).¹¹

College	October 2021 - June 2022	October 2022 - June 2023
A	0%	5%
B	14%	0%
C	5%	5%
D	28%	28%

Table 1: College subsidies on vegetarian lunches compared to meat alternative

The origins of these pricing decisions support treating them as unrelated to student dietary trends. At College A, the subsidy was introduced after a small group of environmentally-motivated students proposed it to college committees, intended to signal the environmental impact of food choices and encourage vegetarian consumption among the broader student population. At College B, the withdrawal reflected catering cost considerations. As noted by the catering manager, student meals are subsidised across the board, and vegetarian and low-carbon ingredients are often no cheaper than meat at a comparable quality. Neither change was a response to shifting vegetarian demand, consistent with the parallel pre-trends documented in Appendix C.

2.3 Data

We limit our analysis to cafeteria lunch and dinner main meals only, excluding extra sides, salads, soups and desserts, as well as any breakfast items or special dinners. The mains are the focus of our study as at each mealtime the colleges always offer mains options in both the vegetarian and non-vegetarian categories.

For one of our colleges, the meal type coding is limited to vegan/vegetarian or meat/ fish. We therefore focus our analysis on comparing vegetarian (including

¹¹The subsidies in Table 1 are expressed as whole-number percentages. Exact subsidies varied slightly around the whole number across years as cafeterias adjusted prices to maintain whole-number prices.

vegan) and non-vegetarian options (fish and meat). Our ‘full dataset’ spanning both academic years (October 2021 - June 2023) includes 186,851 meal choices.

Checking for general consistency of sales and similarity of behaviour across terms and colleges Figure A.1 (Appendix C) plots weekly lunch sales of the full 2-year dataset and monthly vegetarian meal sales data per college, respectively. Treatment colleges collectively enrol about 8% more students than the control colleges. The figure also exhibits the expected term-time peaks and between-term troughs in sales. The timing and shapes of in-term sales are similar for treatment and control colleges.

For heterogeneity analysis, students are partitioned into groups according to their baseline dietary preferences. Students are grouped based on whether pre-subsidy introduction or removal they consumed predominantly vegetarian meals ($> 70\%$) these are our ‘most vegetarian’ consumers, very few vegetarian meals ($< 30\%$), our ‘least vegetarian’ consumers, or somewhere in-between those two levels (30-70%), our ‘flexitarian’ cohort.¹²

2.4 Dependent and Other Variables

The dependent variable is a mealtime-choice indicator equal to one if a vegetarian option is chosen and zero otherwise. This coding follows common practice and facilitates comparison with the existing literature (Lohmann et al., 2022; Garnett et al., 2021, 2019; Kurz, 2018). Although not all vegetarian meals are strictly lower-carbon than all non-vegetarian meals, on average, animal-based foods have higher environmental impacts than vegetarian diets (Tilman and Clark, 2014; Clark et al., 2022; Scarborough et al., 2023). Lohmann et al. (2022) finds that vegetarian meals have lower average carbon footprints than meat-based meals in a comparable

¹²This grouping departs from the quartile split registered in our pre-analysis plan; we document this and our reasons in Appendix A.1. We also report heterogeneity analysis using a continuous measure of baseline vegetarian share (Appendix Tables A.2 and A.3).

college cafeteria setting.

We control for weekday-specific effects at the college level by interacting day-of-week indicators with college fixed effects, capturing specific menu patterns (for example, a recurring Monday menu at a given college). We also include an indicator for whether the meal purchase occurs at lunchtime or dinner.

As availability has been shown to affect meal choices (Garnett et al., 2019; Pechey et al., 2022b,a), for each mealtime in each college, we count when the vegetarian, or non-vegetarian, option has not been purchased in the last 30 minutes of a service with more than 15 meal sales in those last 30 minutes. In only 2.8% of cases over the 2-year period (including lunches and dinners) does this occur. Cafeteria chefs indicated that each college followed a policy of replenishing meal types in response to demand. This variable is endogenous to the dependent variable so cannot be included as a control in the regression.

2.5 Research Questions

Our design enables three sets of comparisons, each motivated by competing theoretical accounts of how pricing policies shape behaviour.

Average effects. Standard price theory predicts that a subsidy increases demand for the subsidised good; removing one decreases it. However, if subsidies crowd out intrinsic motivation for pro-social choices, the introduction effect may be attenuated or even negative, and withdrawal need not restore intrinsic demand if the crowding-out effect is long lasting (Perino et al., 2014; Bowles and Polania-Reyes, 2012; Bénabou and Tirole, 2006; Cassar and Meier, 2021; Fehr and Fischbacher, 2002). We estimate the average treatment effect of introducing a 5% subsidy (College A vs. College C and D) and, separately, of withdrawing a 14% subsidy (College B vs. College C and D).

Heterogeneity by baseline diet. Price responsiveness should be concentrated among consumers near the margin of switching — those for whom the utility cost of choosing vegetarian is close to the price differential. We partition students by pre-intervention vegetarian meal share ($>70\%$, $30\text{--}70\%$, $<30\%$) and estimate group-specific treatment effects. If responses are driven by marginal consumers rather than population-wide preference shifts, effects should be largest among those already consuming higher amounts of vegetarian meals pre-intervention.

Dynamics and persistence. Endogenous preference and habit-formation models predict that sustained exposure to the subsidised behaviour may produce durable change that outlasts the policy (Becker and Murphy, 1988; Holt and Goodwin, 1997; Bowles and Polania-Reyes, 2012; Mattauch et al., 2022). Salience-based accounts instead predict an immediate response that may fade as the new price regime becomes the reference point (Bordalo et al., 2013, 2020). Standard price responsiveness predicts stable effects that are contingent on the policy remaining in place. The withdrawal experiment provides a direct test: if habits formed during the subsidy period, removal should have little effect. We use the withdrawal ATE together with week-level event studies for both introduction and withdrawal to distinguish these theories.

2.6 Estimation Strategy

The estimation strategy is in two parts, estimating the impact of introducing a subsidy (College A) and, separately, the impact of removing a subsidy (College B), compared to control colleges that did not change their pricing policies during the study period (Colleges C and D). For both, we implement the same specifications and robustness checks. We begin with a two-way fixed effects linear difference-in-difference (DiD) design with person and time fixed effects, absorbing time-invariant

individual heterogeneity and common time shocks (Bertrand et al., 2004).¹³ Our baseline estimation regression is defined below.

Let i index individuals, s meal services (college $\times$ mealtime), and t the date. The outcome variable $\text{Veg}_{i,s,t}$ equals to 1 if individual i selects a vegetarian/vegan main dish at service s on day t , and 0 for a meat or fish one. Our generalised equation follows:

$$\text{Veg}_{i,t,s} = \alpha_i + \lambda_t + \delta^{DD} D_{i,s} + X_s + \epsilon_{i,t,s} \quad (1)$$

α_i and λ_t capture the individual and week fixed effects, respectively. $D_{i,s} \equiv \text{Treat}_c \times \text{Post}_{c,t}$. Treat_c indicates with 1 if the cafeteria is the treatment college for the specific research question, 0 for its control. The treatment college not used for a research question is excluded from the dataset. $\text{Post}_{c,t}$ indicates with 1 whether the specific meal-service takes place the price interventions were applied (added or removed) in the treatment colleges, or 0 if it was during the baseline year. X_s is a vector of time-varying service-level controls (day-of-week dummies by college, unique week dummies and, a lunch vs. dinner indicator). The error term $\epsilon_{i,t,s}$ will be clustered at different levels, as specified in the inference section discussion.

We use equation (1) separately for both streams of research questions, adding a subsidy and discontinuing one. First, we study subsidy introduction using data from the college that adopts a subsidy in the second year (College A) and the control colleges that always apply a subsidy (College C and D) for any background unmeasured effects. Second, we examine subsidy withdrawal effects using the college that removes a subsidy in the second year (College B) and the same control colleges. For both of these, our main analysis is based on two academic years of

¹³Controlling for individual fixed effects is crucial when studying food consumption habits. Carasco et al. (2005) show that failing to account for household fixed effects conflates unobserved heterogeneity with true habit formation in food demand, yielding misleading conclusions about intertemporal dynamics.

data, 2021-22 and 2022-23. For the research question on the effects of discontinuing the subsidy, the mealtime used is only lunchtimes, as the subsidy was only applied to lunch meals in College B.

We use a week-level event study analysis to capture dynamic treatment effects over time from the point of intervention being studied, as done by Schulze-Tilling (2025). Finally, we examine heterogeneous effects of the price intervention by consumption types. We classify individuals by their baseline (pre-intervention) vegetarian-meal consumption and estimate group-specific DiD coefficients by interacting the treatment indicator with group dummies.

For our primary results, we report standard errors clustered at the individual level. As a robustness check (Table A.1), we also compute standard errors clustered at the college level using the wild-cluster bootstrap, as well as two-way clustering by college and week. However, with only one treated and two control colleges, these cluster-robust methods do not satisfy their asymptotic requirements (Cameron et al., 2008, 2011; Roodman et al., 2019; Abadie et al., 2023; MacKinnon and Webb, 2018). We test for residual heterogeneity. We find no evidence of systematic variation between colleges in the residuals from our main regressions. However, we do find evidence of serial correlation in residuals within individuals and shared residual shocks pre- and post-intervention by individuals within the same college. To address these, we include individual and college-by-week standard-errors to capture those sources of heterogeneity, and cluster standard errors at the individual level to account for serial correlation. Roth et al. (2023) show through model-based assumptions that it can be permissible to cluster at a more granular-level than the treatment level in DiD settings with few clusters. The authors suggest that this is possible if the cluster shocks are small relative to idiosyncratic variation.¹⁴

¹⁴In their paper, they cite the example of Card and Krueger (2000) which studies unemployment in two cities, a case like ours where clustering at the city level is not possible.

Identification in the DiD framework relies on the assumption of parallel trends in the absence of treatment. To evaluate this assumption, Figure A.1 plots regression-adjusted monthly vegetarian uptake rates for colleges during the pre-intervention period (2021–22). While levels vary modestly across colleges, the trends are stable and largely parallel throughout the baseline year.

3 Results

3.1 Descriptive Statistics

3.1.1 Sample composition

Table 2 provides an overview of the sample composition for our two-year fixed effects regressions. In the two-year dataset, which includes only students present in both Year 1 and Year 2, there are approximately 54,000 and 50,000 total sales in the treatment and control colleges, respectively. The sample comprises 668 students in the treatment colleges and 631 in the control colleges, with an average of 43 and 50 meals purchased per person, respectively, during the baseline year.

Table 2: Treatment and control sample statistics over the study periods

	Treatment Colleges		Control Colleges	
	Baseline	Intervention	Baseline	Intervention
First two years and people in both only, FE regressions				
Days	184	185	184	189
Individuals	668	668	631	631
Avg meals per student and per year	43.5	37.4	44.8	34.9
Individual sales	28,767	24,742	28,062	21,832
Total sales	N = 53,509		N = 49,894	

Note: Table provides an overview of the analysis sample using all available meal choice observations during term time. Baseline = 2021-22; intervention = 2022-23. "Treatment colleges" combines Colleges A and B. "Control colleges" combines Colleges C and D.

3.1.2 General Trends in Food Consumption Preferences

Before turning to the treatment effects, we document the broader backdrop against which the pricing interventions took place. Both control colleges show a declining trend in vegetarian meal uptake over the study period, and treatment colleges follow a similar baseline trajectory. This common downward drift is absorbed by the time fixed effects in our difference-in-differences specification, and pre-trend evidence (Appendix C) supports the parallel-trends assumption. Importantly, the decline is present within the balanced panel of continuing students who identify our fixed-effects estimates, and is concentrated among consumers with higher-baseline consumption of meat, this is common to the treatment and control colleges. Appendix C.2 further shows that successive fresher cohorts enter with progressively lower vegetarian shares, confirming that the aggregate trend is not because of changing sample composition. Any subsidy-induced gains must thus be interpreted against the backdrop of this decrease in vegetarian demand.

3.2 Adding a Vegetarian Meal Subsidy

We first examine the average treatment effect (ATE) associated with the introduction of a 5% subsidy on vegetarian meals (Table 3). Estimates are drawn from equation (1), comparing the treated college to the control colleges. Standard errors are clustered at the individual-level (Table 3, columns 1-2) and cafeteria level (column 3).

Across specifications, we find that the introduction of a 5% subsidy increases vegetarian meal consumption by approximately 3 percentage point (pp), a modest but consistent effect that partially offsets the general downward trend in vegetarian demand observed across all colleges.¹⁵ We find a slightly larger effect (3.2 pp

¹⁵Clustering standard errors at the cafeteria level (column 3), we find that our primary treatment effect is still statistically significant, as with the individual-level clustering. Yet with only three clusters, the college-level inference is not asymptotically validated. We include full cafeteria-level clustered error terms with a comparison to inference with the wild bootstrap and the sub-cluster

increase) when controlling for baseline dietary preferences (and significant triple interaction of 17.5 pp, at the 1% level)(Table A.2) and 3.5 pp when considering a trimmed shorter spring term (Table A.7).¹⁶

Table 3: Main Results Adding 5% Subsidy

	(1)	(2)	(3)
Adding 5% Discount $\times$ Treat	0.029** (0.012)	0.029** (0.012)	0.029*** (0.001) [0.0440]
Dinner		0.017*** (0.004)	0.017 (0.024)
Constant	0.357*** (0.002)	0.348*** (0.003)	0.348*** (0.012)
Cafeteria $\times$ Day-of-Week FEs	yes	yes	yes
Student FEs	yes	yes	yes
Week FEs	yes	yes	yes
Cafeteria FEs	yes	yes	yes
Clusters	Student Level	Student Level	Cafeteria Level
Students	990	990	990
Meals	89114	89114	89114
R-squared (within)	0.0003	0.0007	0.0007

Clustered-standard errors in parentheses. Lunch mealtime is reference of dinner.

Col. (3): wild cluster bootstrap p-value (Webb weights, 999 reps.) in square brackets. With only 3 cafeterias, cafeteria bootstrap under-rejects: indicative only (MacKinnon and Webb, 2018).

OLS Difference-in-difference regression based on first 2 years of data.

Sample includes College A (treated), Colleges C and D (controls).

Week effects are unique to each year. Day-of-week effects are unique to each cafeteria.

Within-R2 values exclude explanatory value from FE variables

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

inference in the Appendix A.2 Table A.1.

¹⁶As a further robustness check, we re-estimate the main results excluding the final three weeks of spring term in 2021–22 and 2022–23, when exam schedules and end-of-year social events may alter cafeteria usage. The results are consistent with the main analysis (Appendix B).

3.3 Removing a Vegetarian Meal Subsidy

Next, we examine the average treatment effect of withdrawing a 14% vegetarian meal subsidy using the same fixed-effects specification, comparing the college that removed its subsidy at the start of the second year (College B) to the colleges that maintained their subsidies throughout (College C and D).¹⁷ This analysis is limited to lunchtime sales only, as College B did not apply the subsidy on dinners (see Table A.12).

We find a weakly significant (at the 10% level) average treated effect on the demand for vegetarian meals from removing the subsidy (Table 4). Our baseline specification suggests a 2.7 pp decrease in demand, mirroring the ATE of introducing a subsidy reported in Table 3. The effect is only somewhat robust i) when controlling for baseline dietary preferences, where the decrease is concentrated among students with higher baseline vegetarian shares (only significant triple interaction of -12.9 pp, significant at the 5% level; Table A.3), and ii) larger (-3.7 pp, significant at the 5% level) when using shorter spring terms to check robustness against potential end-of-term effects (Table A.8).¹⁸

¹⁷The subsidy withdrawal analysis was not pre-registered, as the pricing variation only became apparent during data curation, after the pre-registration of the subsidy-introduction analysis. To limit researcher degrees of freedom, we apply exactly the same pre-registered empirical specifications and robustness checks as in Section 3.2, and we report these withdrawal results as exploratory rather than confirmatory.

¹⁸The shorter spring term robustness check excludes the final three weeks of spring term in 2021–22 and 2022–23, when exam schedules and end-of-year social events may alter cafeteria usage as discussed in Appendix B.

Table 4: Main Results Removing 14% Subsidy

	(1)	(2)
Removing 14% Discount $\times$ Treat	-0.027* (0.015)	-0.027 (0.011) [0.3834]
Constant	0.347*** (0.002)	0.347*** (0.002)
Cafeteria $\times$ Day-of-Week FEs	yes	yes
Student FEs	yes	yes
Week FEs	yes	yes
Cafeteria FEs	yes	yes
Clusters	Student Level	Cafeteria Level
Students	855	855
Meals	44224	44224
R-squared (within)	0.0002	0.0002

Clustered-standard errors in parentheses. Lunch mealtime is reference of dinner.

Col. (2): wild cluster bootstrap p-value (Webb weights, 999 reps.) in square brackets. With only 3 cafeterias, cafeteria bootstrap under-rejects: indicative only (MacKinnon and Webb, 2018).

OLS Difference-in-difference regression based on first 2 years of data.

Sample includes College B (treated), Colleges C and D (controls). Only lunches are included.

Week effects are unique to each year. Day-of-week effects are unique to each cafeteria.

Within-R2 values exclude explanatory value from FE variables

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

4 Heterogeneity Analysis: Effects by Underlying Diet Preference

We now examine heterogeneous treatment effects by baseline dietary preference, grouping students into the three categories defined in Section 3.1.2: least vegetarian (<30%), flexitarian (30–70%), and most vegetarian (>70%) meal choices in their baseline year. Tables 5 and 6 report the fixed effects estimates from equation (1) separately for each group, for the subsidy introduction and withdrawal respectively. We assess robustness to the grouping in two ways in the Appendix: a continuous specification interacting treatment with the baseline vegetarian share (Tables A.2 and A.3), and the pre-registered quartile grouping (Tables A.5 and A.6) both of which recover the same pattern (significant effects among the higher baseline groups, and none with the lowest). The heterogeneity analysis is also robust to restricting to shorter spring terms (Tables A.9 and A.10).

Table 5 shows clear heterogeneity in response to the 5% subsidy. The subsidy has no significant effect on students who consumed less than 30% of vegetarian meals in their baseline year, while both the most vegetarian and flexitarian groups increase their vegetarian meal uptake by 6 pp.

Table 5: Heterogeneity Analysis Adding 5% Subsidy

	> 70% Veg	30-70% Veg	< 30% Veg
Adding 5% Discount $\times$ Treat	0.059** (0.025)	0.061** (0.031)	-0.004 (0.013)
Dinner	0.003 (0.008)	0.015* (0.009)	0.022*** (0.005)
Constant	0.847*** (0.006)	0.434*** (0.006)	0.150*** (0.003)
Cafeteria $\times$ Day-of-Week FEs	yes	yes	yes
Student FEs	yes	yes	yes
Week FEs	yes	yes	yes
Cafeteria FEs	yes	yes	yes
Clusters	Student Level	Student Level	Student Level
Students	191	322	477
Meals	13311	31470	44333
R-squared (within)	0.0014	0.0009	0.0008

Clustered-standard errors in parentheses. Lunch mealtime is reference of dinner.

OLS Difference-in-difference regression based on first 2 years of data.

Sample includes College A (treated), Colleges C and D (controls).

Week effects are unique to each year. Day-of-week effects are unique to each cafeteria.

Within-R2 values exclude explanatory value from FE variables

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 6 presents our heterogeneity analysis by baseline dietary preference for the removal of the subsidy. The heterogeneity analysis reveals substantial and statistically significant effects. Notably those with the middle level of vegetarian meal consumption (30–70%), as identified when the subsidy is in place, reduce vegetarian consumption by 6.7 pp after the subsidy is removed, over and above the general downward trend. This is compared to the previously established weakly significant ATE decrease of 2.7 pp (Table 4). The most vegetarian students (those consuming more than 70% vegetarian meals in the baseline year) are estimated to reduce their vegetarian meal uptake by 4.9 pp following its removal, however,

this is not a significant estimate.¹⁹ The least vegetarian group shows no significant change in purchase behaviour in any of the model specifications. As a further robustness check, and to avoid selecting fixed thresholds entirely, we also estimate a specification interacting treatment with the continuous baseline vegetarian share (Table A.3). This recovers the same gradient: responsiveness to withdrawal increases with baseline vegetarian consumption, indicating that the concentration of the effect among higher-baseline consumers does not depend on the chosen cut points.

Table 6: Heterogeneity Analysis Removing 14% Subsidy

	> 70% Veg	30-70% Veg	< 30% Veg
Removing 14% Subsidy × Treat	-0.049 (0.044)	-0.067*** (0.026)	-0.002 (0.016)
Constant	0.850*** (0.008)	0.442*** (0.004)	0.158*** (0.003)
Cafeteria × Day-of-Week FEs	yes	yes	yes
Student FEs	yes	yes	yes
Week FEs	yes	yes	yes
Cafeteria FEs	yes	yes	yes
Clusters	Student Level	Student Level	Student Level
Students	153	261	441
Meals	5558	17565	21101
R-squared (within)	0.0010	0.0009	0.0000

Clustered-standard errors in parentheses.

OLS Difference-in-difference regression based on first 2 years of data.

Sample includes College B (treated), Colleges C and D (controls). Only lunches are included.

Week effects are unique to each year. Day-of-week effects are unique to each cafeteria.

Within-R2 values exclude explanatory value from FE variables

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Taken together, the heterogeneity analysis reveals that the average treatment effects

¹⁹We do note the higher standard error estimate with the fewer students compared to the other heterogeneity groups and test a smaller more defined heterogeneity level (those consuming more than 80% vegetarian meals in the baseline year) in hopes of reducing variance. In Table A.4 we find that those consuming more than 80% vegetarian meals in the baseline year reduce their vegetarian meal uptake by 8.2 pp following the subsidy removal, which is significant at the 5% level.

reported in Sections 3.2 and 3.3 mask substantial variation across dietary preference groups. In both the introduction and withdrawal case, it is the most and middle-level vegetarian students who respond most strongly to the pricing change, increasing uptake when the subsidy is introduced and reducing it sharply when it is removed. However, it is important to note that the students in this analysis experienced both the before and after of the pricing change, making the price shift salient in a way that may not generalise to new cohorts who have no prior pricing reference point. The responses we estimate may therefore partly reflect a reaction to a salient price change (Bordalo et al., 2013, 2020), rather than a durable shift in behaviour. We investigate this further using event-study analysis reported next.

5 Event Study: Intervention spillover

The canonical difference-in-differences (DiD) model estimates the average effect of the price intervention but does not capture its dynamics over time. Consumer responses might be delayed as they gradually notice the price change, fade as the new pricing becomes less salient (Bordalo et al., 2013, 2020), or strengthen if the policy encourages habit formation or shifts underlying preferences (Becker and Murphy, 1988; Holt and Goodwin, 1997; Bowles and Polania-Reyes, 2012; Mattauch et al., 2022). To explore these possibilities, we complement the DiD estimates with a month-level event study.

We present event-study estimates separately for the most vegetarian two groups combined ($>70\%$ and $30-70\%$) and least vegetarian ($<30\%$) groups. We combine the two most vegetarian groups, which our heterogeneity analysis suggests are similarly responsive to the interventions, and helps maintain a sufficient sample size for statistical power while reducing the study intervals from full sample to monthly. The reference point is the monthly estimate of June 2022 (the end of spring term 2021–22). Note that estimates are not comparable across groups, as

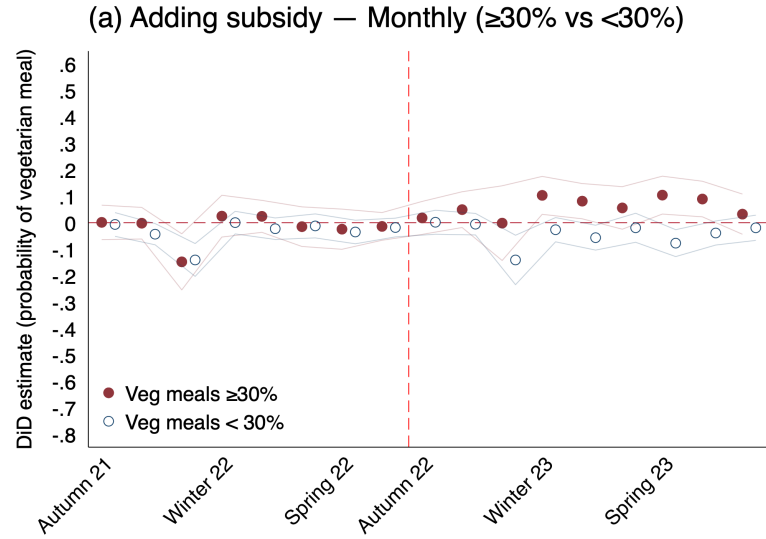
the pre-intervention reference is defined within each preference group.

Figure 1a shows the weekly effects of adding the subsidy in College A by preference group. Among the two most vegetarian students ($\geq 30\%$), there is an upward shift in vegetarian meal consumption following the introduction of the 5% subsidy in Autumn 2022, consistent with the estimates in Table 5. The effect appears within the first few months and shows no sign of decay over the academic year, other than perhaps a slight downward trends towards the end of each term. The least vegetarian group shows no significant response.

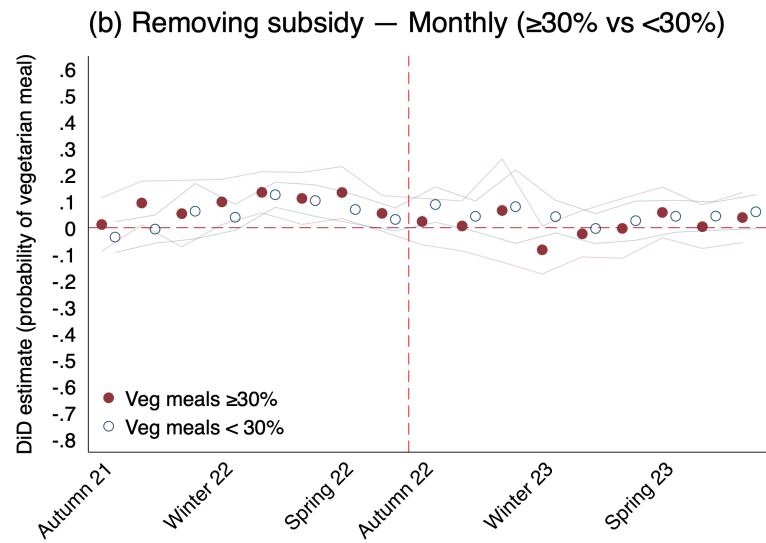
Figure 1b highlights some contrast between the two groups following the subsidy removal. While the least vegetarian group ($>30\%$) remains broadly stable before and after the intervention in Autumn 2022, the two most vegetarian group's estimates drop markedly and immediately at the start of the second year. This negative shift appears stable throughout Year 2.

6 A Note on Cost-effectiveness of subsidy introduction

A holistic assessment of the efficacy of vegetarian meal subsidies as a policy tool must also consider their ability to translate into meaningful greenhouse gas (GHG) emissions reductions and their corresponding cost-effectiveness. Following Schulze-Tilling (2025), we conduct a simple back-of-the-envelope calculation using GHG emissions data of a comparable sample of cafeteria meals (Lohmann et al., 2022) to complement our empirical results. An average vegetarian or vegan meal has a GHG emissions intensity of 0.8 kg CO₂e, whereas an average non-vegetarian meal has a footprint of 2.5 kg CO₂e, resulting in a difference of 1.7 kg CO₂e. Based on the proportions of meat/fish (68%) and vegetarian/vegan (32%) meals in the treatment cafeteria prior to the intervention, the weighted average baseline emis-



(a) Adding subsidy



(b) Removing subsidy

Figure 1: Event-study DiD estimates

Notes. Estimates from a monthly event-study extension of equation (1), estimated separately by baseline preference group. 95% CIs shown as shaded areas. Baseline = Year 1, June 2022. Vertical dashed red line marks the start of the second academic year.

sions were approximately 1.96 kg CO₂e per meal.

The intervention caused a 3 pp shift from meat meals to vegetarian meals, as estimated from our primary analysis. Given the emissions difference between meat (2.5 kg CO₂e) and vegetarian meals (0.8 kg CO₂e), this shift translates to an emissions reduction of 0.051 kg CO₂e (51 grams) per average meal. This represents a 2.65% reduction relative to the baseline emissions of 1.956 kg CO₂e per meal. Scaling this reduction to the total number of meals sold in the treatment cafeteria during the first year of intervention (30,939 meals), the subsidy resulted in an estimated total emissions reduction of 1,577.89 kg CO₂e. Extrapolating this effect over two years to approximately 100,000 meals, assuming consistent intervention impact, the intervention would lead to an additional reduction of 5,100 kg CO₂e.

The subsidy introduced a relative discount of 5%, making vegetarian meals £0.15 cheaper at lunch and £0.20 cheaper at dinner. During the study period, 4,707 vegetarian meals were sold at lunch and 5,774 at dinner, resulting in a total intervention cost of £1,860.85. With emissions savings of approximately 1,577.89 kg CO₂e, the cost-effectiveness of the intervention is therefore estimated at £1.18 per kg CO₂e saved, or £1,180 per tonne CO₂e saved.

When compared to meta-analytic estimates of the social cost of carbon (SCC) from Moore et al. (2024), updated to 2025 values of £42.31 (median) and £143.08 (mean) per tonne CO₂e, the subsidy's abatement cost of approximately £1,180 per tonne CO₂e is substantially higher – by factors of 27.9 and 8.2, respectively. Moreover, this figure greatly exceeds the illustrative abatement cost of around £31 per tonne CO₂e estimated for a cafeteria carbon-labelling intervention under comparable assumptions Lohmann et al. (2022). These comparisons suggest that, while the intervention successfully reduced emissions, its financial cost relative to the societal value of emissions reductions may limit its scalability in practice. It further points to the need of careful targeting towards the lowest emitting options, such as

exclusively subsidising plant-based dishes as opposed to all vegetarian meals (Pizzo et al., 2026).

These back-of-the-envelope calculations underscore the significant inefficiency of environmental subsidies in this specific context, particularly when contrasted with the theoretical efficiency of tax-based instruments in addressing environmental externalities. Nonetheless, in political discourse and among the broader public, carbon taxes on meat and dairy products generally face greater opposition than subsidies for more sustainable consumption, which tend to enjoy higher levels of public acceptability (Pechey et al., 2022c). For instance, Lohmann et al. (2024) report that only 15% of UK citizens oppose vegetarian subsidies, compared to 51% who oppose a meat tax.

7 Discussion and Conclusion

The importance of pricing policy studies is not only whether incentives change behaviour at the point of intervention, but whether those changes persist once the incentive is removed (Royer et al., 2015; Carrera et al., 2020; Charness and Gneezy, 2009). Using two natural experiments in university cafeterias, this paper studies both questions: the introduction of a small subsidy on vegetarian meals and the withdrawal of a pre-existing subsidy.

We find that consumers respond to price incentives, but only selectively and only while those incentives remain in place. Introducing a 5% subsidy increases vegetarian meal choices on average, but the effect is concentrated among students who were already more vegetarian in their baseline behaviour. Removing a 14% subsidy mirrors the average effect of the inclusion of the subsidy in the opposite direction, yet leads to heterogeneous effects among different (baseline year) levels of vegetarian consumers. Taken together, these results are more consistent with heterogeneous price responsiveness than with broad, durable changes in underlying dietary preferences (Becker and Murphy, 1988; Houde and Aldy, 2017; Charness and Gneezy, 2009).

This interpretation is reinforced by our event-study analysis which suggests that responses occur quickly around the policy change and then remain broadly tied to the policy environment rather than evolving into a lasting behavioural shift. In this setting, subsidies appear to move behaviour rapidly, but not to generate strong evidence of internalised or social habit formation or persistent preference change after withdrawal (Becker and Murphy, 1988; Mattauch et al., 2022). Because we observe both subsidy introduction and subsidy withdrawal in comparable institutional settings with repeated individual-level transactions over multiple years, the data allow a test of persistence that most single-policy, short-horizon studies cannot provide (Carrera et al., 2020; Charness and Gneezy, 2009; Royer et al., 2015;

Garnett et al., 2019). One channel through which persistent effects could plausibly emerge is experiential learning. Klatt and Schulze Tilling (2024) find lasting increases in vegetarian demand following a vegetarian month, because temporarily inducing diners to try vegetarian meals led them to update their beliefs about options they would not otherwise have chosen. In our setting, however, the students who would stand to gain most from this effect, those with low baseline vegetarian demand, are precisely the group that does not change behaviour when the subsidy is introduced, leaving little scope for such preference updating to take hold.

Prior work documents that price changes can shift vegetarian demand over short horizons (Garnett et al., 2019). Our results show that comparable average effects persist over a year-long horizon, but also that the gains from subsidy introduction are not uniform across our study population and are largely unwound for the most responsive consumers once the subsidy is removed.

We find that treatment effects are highly heterogeneous. The subsidy mainly shifts behaviour among students already near the margin of choosing vegetarian meals, while students with persistently low baseline vegetarian demand respond little to the incentives. This limits the aggregate impact of a small non-targeted policy, and highlights the importance of accounting for differences in responsiveness across consumers in the design of corrective subsidies (Kiso, 2026). In our setting, those most affected by the policy are not those with the greatest potential for emissions reduction savings from changing meal choices. More broadly, the results show that average treatment effects alone can mask economically important differences in policy incidence across consumers.

The cost-effectiveness calculations reinforce this point (see Appendix Section 6 for details). Although the subsidy produces measurable emissions savings, the implied abatement cost of £1,180 per tonne CO₂e remains high relative to standard social-cost-of-carbon benchmarks and available estimates for alternative cafeteria inter-

ventions such as carbon labelling (Lohmann et al., 2022; Schulze-Tilling, 2025; Moore et al., 2024). Taken together with the heterogeneity results, this suggests that modest vegetarian subsidies may be difficult to justify as a stand-alone climate policy in cafeteria settings.

Our findings should also be interpreted with several limitations in mind. First, Because introduction and withdrawal occur in different colleges, we do not observe the same students gain and then lose a subsidy, and so cannot test persistence as a within-population reversal. Instead, we observe two populations responding in mirror-image fashion, with effects concentrated in the same high-baseline consumers and, in each college, an immediate response that remains tied to the policy while it is in place. This parallel pattern, rather than a within-person reversal, underpins our interpretation of the results as heterogeneous price responsiveness rather than durable preference change. Second, the analysis is based on university cafeteria choices, and the external validity of these estimates for supermarkets or other food environments is uncertain. Third, while we observe repeated within-cafeteria meal choices, we do not observe substitution outside the cafeteria or total diet, so the estimated effects should be interpreted as within-setting behavioural responses. Relatedly, each academic year comprises three 8-week terms separated by vacation periods of roughly 6 weeks during which students leave college accommodation and return, presumably, to their homes. We have no information on what dietary choices dominate during those intervals, so we cannot rule out that home food environments partly drive or dilute the patterns we observe. On the upside, because the in-term effects we document must survive repeated gaps in cafeteria exposure, the within-year stability we report is a more demanding form of persistence than continuous-exposure designs can offer. Behaviour that re-emerges at the start of each term, after a break from the subsidised context, provides at least some evidence against a purely salience-driven or attention-based response that would fade quickly without reinforcement. Fourth, although the heterogeneity patterns are

clear, we cannot separately identify whether they arise from differences in tastes, salience, budget sensitivity, identity, or other possible mechanisms.

Overall, the evidence suggests that modest food-price subsidies can change behaviour, but mainly for consumers already near the margin and only for as long as the incentive remains in place. Because the effect fades once the subsidy is withdrawn, sustaining it would mean funding the subsidy indefinitely, at a cost well above standard abatement benchmarks. Durably shifting the diets of habitual meat-eaters, who account for over 80% of the UK population, is therefore unlikely to be achieved by price incentives alone, and points to the need for instruments that can shift habits rather than prices that must be maintained to have any effect.

References

- Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. (2023). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, 138(1):1–35.
- Acemoglu, D., Aghion, P., Bursztyn, L., and Hémous, D. (2012). The environment and directed technical change. *American Economic Review*, 102(1):131–166.
- An, R. (2013). Effectiveness of subsidies in promoting healthy food purchases and consumption: a review of field experiments. *Public Health Nutrition*, 16(7):1215–1228.
- Andor, M. A., Dehos, F., Gillingham, K., Hansteen, S., and Tomberg, L. (2023). Promoting public transport with modern pricing schemes. Technical Report 82, RWI – Leibniz-Institut für Wirtschaftsforschung, Essen.
- Bartels, L. and Werthschulte, M. (2025). ‘more bang for the buck’? experimental evidence on the mechanisms of an energy efficiency subsidy: L. bartels, m. werthschulte. *Environmental and Resource Economics*, 88(3):631–654.
- Becker, G. S. and Murphy, K. M. (1988). A theory of rational addiction. *Journal of Political Economy*, 96(4):675–700.
- Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, 96(5):1652–1678.
- Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, 119(1):249–275.
- Bordalo, P., Gennaioli, N., and Shleifer, A. (2013). Salience and consumer choice. *Journal of Political Economy*, 121(5):803–843.
- Bordalo, P., Gennaioli, N., and Shleifer, A. (2020). Memory, attention, and choice. *The Quarterly Journal of Economics*, 135(3):1399–1442.
- Bowles, S. and Polania-Reyes, S. (2012). Economic incentives and social preferences: substitutes or complements? *Journal of Economic Literature*, 50(2):368–425.
- Bull, O., Muñoz, J. C., and Silva, H. E. (2021). The impact of fare-free public transport on travel behavior: Evidence from a randomized controlled trial. *Regional Science and Urban Economics*, 86:103616.
- Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427.
- Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, 29(2):238–249.
- Card, D. and Krueger, A. B. (2000). Minimum wages and employment: a case

- study of the fast-food industry in new jersey and pennsylvania: reply. *American Economic Review*, 90(5):1397–1420.
- Carrasco, R., Labeaga, J. M., and David López-Salido, J. (2005). Consumption and habits: evidence from panel data. *The Economic Journal*, 115(500):144–165.
- Carrera, M., Royer, H., Stehr, M., and Sydnor, J. (2020). The structure of health incentives: Evidence from a field experiment. *Management Science*, 66(5):1890–1908.
- Cassar, L. and Meier, S. (2021). Intentions for doing good matter for doing well: The negative effects of prosocial incentives. *The Economic Journal*, 131(637):1988–2017.
- Charness, G. and Gneezy, U. (2009). Incentives to exercise. *Econometrica*, 77(3):909–931.
- Chetty, R., Looney, A., and Kroft, K. (2009). Salience and taxation: Theory and evidence. *American Economic Review*, 99(4):1145–1177.
- Clark, M., Springmann, M., Rayner, M., Scarborough, P., Hill, J., Tilman, D., Macdiarmid, J. I., Fanzo, J., Bandy, L., and Harrington, R. A. (2022). Estimating the environmental impacts of 57,000 food products. *Proceedings of the National Academy of Sciences*, 119(33):e2120584119.
- Clark, M. A., Domingo, N. G., Colgan, K., Thakrar, S. K., Tilman, D., Lynch, J., Azevedo, I. L., and Hill, J. D. (2020). Global food system emissions could preclude achieving the 1.5 and 2 c climate change targets. *Science*, 370(6517):705–708.
- Clinton, B. C. and Steinberg, D. C. (2019). Providing the spark: Impact of financial incentives on battery electric vehicle adoption. *Journal of Environmental Economics and Management*, 98:102255.
- Colmer, J., Martin, R., Muûls, M., and Wagner, U. J. (2025). Does pricing carbon mitigate climate change? firm-level evidence from the european union emissions trading system. *Review of Economic Studies*, 92(3):1625–1660.
- Creutzig, F., Niamir, L., Bai, X., Callaghan, M., Cullen, J., Díaz-José, J., Figueroa, M., Grubler, A., Lamb, W. F., Leip, A., et al. (2022). Demand-side solutions to climate change mitigation consistent with high levels of well-being. *Nature Climate Change*, 12(1):36–46.
- Dannenberg, A., Klatt, C., and Weingärtner, E. (2024). The effects of social norms and observability on food choice. *Food Policy*, 125.
- Dannenberg, A. and Weingärtner, E. (2023). The effects of observability and an information nudge on food choice. *Journal of Environmental Economics and Management*, 120.
- Dechezleprêtre, A., Nachtigall, D., and Venmans, F. (2023). The joint impact of the european union emissions trading system on carbon emissions and economic performance. *Journal of Environmental Economics and Management*, 118:102758.

- DellaVigna, S., Lindner, A., Reizer, B., and Schmieder, J. F. (2017). Reference-dependent job search: Evidence from Hungary. *The Quarterly Journal of Economics*, 132(4):1969–2018.
- Douenne, T. and Fabre, A. (2022). Yellow vests, pessimistic beliefs, and carbon tax aversion. *American Economic Journal: Economic Policy*, 14(1):81–110.
- Fageda, X. and Teixidó, J. J. (2022). Pricing carbon in the aviation sector: Evidence from the European emissions trading system. *Journal of Environmental Economics and Management*, 111:102591.
- Fehr, E. and Fischbacher, U. (2002). Why social preferences matter—the impact of non-selfish motives on competition, cooperation and incentives. *The Economic Journal*, 112(478):C1–C33.
- Fosgaard, T. R., Pizzo, A., and Sadoff, S. (2024). Do People Respond to the Climate Impact of their Behavior? The Effect of Carbon Footprint Information on Grocery Purchases. *Environmental and Resource Economics*.
- Funke, F., Mattauch, L., Bijgaart, I. v. d., Godfray, H. C. J., Hepburn, C., Klenert, D., Springmann, M., and Treich, N. (2022). Toward optimal meat pricing: Is it time to tax meat consumption? *Review of Environmental Economics and Policy*, 16(2):219–240.
- Garnett, E. E., Balmford, A., Marteau, T. M., Pilling, M. A., and Sandbrook, C. (2021). Price of change: Does a small alteration to the price of meat and vegetarian options affect their sales? *Journal of Environmental Psychology*, 75:101589.
- Garnett, E. E., Balmford, A., Sandbrook, C., Pilling, M. A., and Marteau, T. M. (2019). Impact of increasing vegetarian availability on meal selection and sales in cafeterias. *Proceedings of the National Academy of Sciences*, 116(42):20923–20929.
- Gneezy, U., Meier, S., and Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives*, 25(4):191–210.
- Gravert, C. and Collentine, L. O. (2021). When nudges aren't enough: Norms, incentives and habit formation in public transport usage. *Journal of Economic Behavior & Organization*, 190:1–14.
- Griffith, R. (2022). Obesity, poverty and public policy. *The Economic Journal*, 132(644):1235–1258.
- Gsottbauer, E. and van den Bergh, J. C. (2025). Pricing instruments in environmental and climate policy when polluters are boundedly rational. *npj Climate Action*, 4(1):96.
- Holt, M. T. and Goodwin, B. K. (1997). Generalized habit formation in an inverse almost ideal demand system: An application to meat expenditures in the US. *Empirical Economics*, 22(2):293–320.
- Houde, S. and Aldy, J. E. (2017). The efficiency consequences of heterogeneous

- behavioral responses to energy fiscal policies. Technical report, National Bureau of Economic Research.
- Ivanova, D., Barrett, J., Wiedenhofer, D., Macura, B., Callaghan, M., and Creutzig, F. (2020). Quantifying the potential for climate change mitigation of consumption options. *Environmental Research Letters*, 15(9):093001.
- Jayachandran, S., De Laat, J., Audy, R., Pagiola, S., and Sedano Santamaria, F. (2018). Evaluating the permanence of forest conservation following the end of payments for environmental services in uganda. *World Bank Group*.
- Kehlbacher, A., Tiffin, R., Briggs, A., Berners-Lee, M., and Scarborough, P. (2016). The distributional and nutritional impacts and mitigation potential of emission-based food taxes in the uk. *Climatic Change*, 137(1):121–141.
- Kemigisha, E., Babweteera, F., Mugisha, J., and Angelsen, A. (2023). Payment for environmental services to reduce deforestation: Do the positive effects last? *Ecological economics*, 209:107840.
- Kiso, T. (2026). Cost-effective and equitable targeting of corrective subsidies: Differentiated green subsidies for new and existing homes. *Journal of Environmental Economics and Management*, page 103340.
- Klatt, C. and Schulze Tilling, A. (2024). Tastes better than expected: post-intervention effects of a vegetarian month in the student canteen. Technical report, ECONtribute Discussion Paper.
- Kőszegi, B. and Rabin, M. (2006). A model of reference-dependent preferences. *The Quarterly Journal of Economics*, 121(4):1133–1165.
- Kurz, V. (2018). Nudging to reduce meat consumption: Immediate and persistent effects of an intervention at a university restaurant. *Journal of Environmental Economics and Management*, 90:317–341.
- Lehmann, J., Truebner, M., Patzina, A., Jeitler, M., Hoffmann, R., and Kessler, C. S. (2025). The willingness to transition to a more plant-based diet among omnivores: Determinants and socioeconomic differences. *Appetite*, 206:107765.
- Lohmann, P. M., Gsottbauer, E., Doherty, A., and Kontoleon, A. (2022). Do carbon footprint labels promote climatarian diets? evidence from a large-scale field experiment. *Journal of Environmental Economics and Management*, 114:102693.
- Lohmann, P. M., Gsottbauer, E., Farrington, J., Human, S., and Reisch, L. A. (2024). Choice architecture promotes sustainable choices in online food-delivery apps. *PNAS Nexus*, 3(10):pgae422.
- Lohmann, P. M., Pizzo, A., Bauer, J. M., Khanna, T. M., Flecke, S. L., Callaghan, M., Minx, J. C., and Reisch, L. A. (2026). A meta-analysis assessing the effectiveness of demand-side interventions for sustainable food consumption and food waste reduction. *Nature Food*, pages 1–12.
- MacKinnon, J. G. and Webb, M. D. (2018). The wild bootstrap for few (treated) clusters. *The Econometrics Journal*, 21(2):114–135.

- Mattauch, L., Hepburn, C., Spuler, F., and Stern, N. (2022). The economics of climate change with endogenous preferences. *Resource and Energy Economics*, 69:101312.
- Metcalf, G. E. (2021). Carbon taxes in theory and practice. *Annual Review of Resource Economics*, 13(1):245–265.
- Moore, F. C., Drupp, M. A., Rising, J., Dietz, S., Rudik, I., and Wagner, G. (2024). Synthesis of evidence yields high social cost of carbon due to structural model variation and uncertainties. *Proceedings of the National Academy of Sciences*, 121(52):e2410733121.
- Mytton, O. T., Clarke, D., and Rayner, M. (2012). Taxing unhealthy food and drinks to improve health. *Bmj*, 344.
- Panzone, L. A., Auch, N., and Zizzo, D. J. (2024). Nudging the food basket green: the effects of commitment and badges on the carbon footprint of food shopping. *Environmental and Resource Economics*, 87(1):89–133.
- Panzone, L. A., Ulph, A., Hilton, D., Gortemaker, I., and Tajudeen, I. A. (2021a). Sustainable by design: choice architecture and the carbon footprint of grocery shopping. *Journal of Public Policy & Marketing*, 40(4):463–486.
- Panzone, L. A., Ulph, A., Zizzo, D. J., Hilton, D., and Clear, A. (2021b). The impact of environmental recall and carbon taxation on the carbon footprint of supermarket shopping. *Journal of Environmental Economics and Management*, 109:102137.
- Parry, I. W. H. and Small, K. A. (2005). Does Britain or the United States have the right gasoline tax? *American Economic Review*, 95(4):1276–1289.
- Pechey, R., Bateman, P., Cook, B., and Jebb, S. A. (2022a). Impact of increasing the relative availability of meat-free options on food selection: two natural field experiments and an online randomised trial. *International Journal of Behavioral Nutrition and Physical Activity*, 19(1):9.
- Pechey, R., Hollands, G. J., and Marteau, T. M. (2022b). Explaining the effect on food selection of altering availability: two experimental studies on the role of relative preferences. *BMC Public Health*, 22(1):868.
- Pechey, R., Reynolds, J. P., Cook, B., Marteau, T. M., and Jebb, S. A. (2022c). Acceptability of policies to reduce consumption of red and processed meat: A population-based survey experiment. *Journal of Environmental Psychology*, 81:101817.
- Perino, G., Panzone, L. A., and Swanson, T. (2014). Motivation crowding in real consumption decisions: Who is messing with my groceries? *Economic Inquiry*, 52(2):592–607.
- Perino, G. and Schwirplies, C. (2022). Meaty arguments and fishy effects: field experimental evidence on the impact of reasons to reduce meat consumption. *Journal of Environmental Economics and Management*, 114:102667.

- Pizzo, A., Bauer, J. M., and Reisch, L. A. (2024). What shapes sustainable food choices?: A field experiment on the impact of a behaviorally informed intervention and a price variation on sustainable food choices. Working paper, Copenhagen Business School.
- Pizzo, A., Bauer, J. M., and Reisch, L. A. (2026). Informational nudges or incentives? a field experiment on vegetarian choices. *Ecological Economics*, 242:108882.
- Roodman, D., Nielsen, M. Ø., MacKinnon, J. G., and Webb, M. D. (2019). Fast and wild: Bootstrap inference in stata using boottest. *The Stata Journal*, 19(1):4–60.
- Roth, J., Sant’Anna, P. H., Bilinski, A., and Poe, J. (2023). What’s trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics*, 235(2):2218–2244.
- Royer, H., Stehr, M., and Sydnor, J. (2015). Incentives, commitments, and habit formation in exercise: evidence from a field experiment with workers at a fortune-500 company. *American Economic Journal: Applied Economics*, 7(3):51–84.
- Scarborough, P., Clark, M., Cobiac, L., Papier, K., Knuppel, A., Lynch, J., Harrington, R., Key, T., and Springmann, M. (2023). Vegans, vegetarians, fish-eaters and meat-eaters in the uk show discrepant environmental impacts. *Nature Food*, 4(7):565–574.
- Schulze-Tilling, A. (2025). The effectiveness of carbon labels. Technical report, University of Bonn and University of Mannheim, Germany.
- Thaler, R. H. (2018). From cashews to nudges: The evolution of behavioral economics. *American Economic Review*, 108(6):1265–1287.
- Tilman, D. and Clark, M. (2014). Global diets link environmental sustainability and human health. *Nature*, 515(7528):518–522.
- Tomasi, M., Faccioli, M., Fezzi, C., and Bateman, I. J. (2026). The welfare impacts of carbon taxes and labels on food demand. *Journal of Environmental Economics and Management*, page 103387.
- van der Ploeg, F. (2025). Why green subsidies are preferred to carbon taxes: Climate policy with heightened carbon tax salience. *Journal of Environmental Economics and Management*, 130(C).

A Appendix: Model Specification Robustness

A.1 Deviations from the Pre-Analysis Plan

The pre-analysis plan (PAP) for the subsidy-introduction study was registered on the Open Science Framework in October 2024, before we accessed and curated the full dataset (https://osf.io/xqg5f/overview?view_only=12af6b486b41406990057eae08574357). Our analysis departs from the registered plan in three respects, described below with our reasoning. Because our design exploits a natural experiment rather than a researcher-implemented one, some analytical choices only became visible during data curation and could not have been fully specified at registration. We pre-registered the analysis before accessing the data as a matter of best practice; where we depart from it below, we give our reasons and assess robustness to each choice.

The subsidy-withdrawal analysis was not pre-registered. At registration we anticipated two colleges applying vegetarian subsidies over the study period. On examining the actual prices in the sales data, we found that one of these had instead removed a pre-existing subsidy. This withdrawal, which forms the second natural experiment, could therefore not have been anticipated at the pre-registration stage, and we report its results as exploratory. We apply the exact same specification, controls, and robustness checks registered for the introduction analysis.

Baseline dietary groups. Following Garnett et al. (2021), the PAP specified quartiles of the baseline vegetarian-meal share. In the paper we instead partition students into three fixed groups: predominantly vegetarian ($> 70\%$), flexitarian ($30 - 70\%$), and predominantly meat ($< 30\%$). We made this change for policy relevance and interpretability. Sample quartiles are defined by the empirical distribution and shift with sample composition, so the same label denotes different underlying diets across samples; the fixed thresholds instead correspond to

recognisable consumer types that are stable across samples and directly meaningful for targeting. Crucially, our heterogeneity conclusions do not depend on this choice. Appendix Section A.5 reports the analysis under the pre-registered quartile grouping, Section A.3 reports specifications interacting the treatment with the *continuous* baseline vegetarian share, for introduction and withdrawal respectively. The quartile split reproduces the pattern from our three-group specification, and the continuous specifications, which involve no grouping or cut points, recover the same gradient. The concentration of effects among higher-baseline consumers is therefore not an artefact of the chosen thresholds.

Level of clustering. The PAP specified cafeteria-level clustering with the wild cluster bootstrap. Our primary specification instead clusters at the individual level, with college-by-week terms. As set out in Section 2.6, with one treated and two control colleges no cafeteria-level cluster-robust method, including the wild cluster bootstrap, meets its asymptotic requirements (Cameron et al., 2008; Roodman et al., 2019; MacKinnon and Webb, 2018). We find no evidence of systematic between-college variation in the residuals, but do find serial correlation within individuals and shared shocks within college-by-week cells; individual-level clustering with college-by-week terms addresses the sources of dependence we actually detect (Roth et al., 2023). We nonetheless report the pre-registered cafeteria-level wild-bootstrap inference as a robustness check (Table A.1). Consistent with the concern that inference with few clusters is fragile, we interpret statistical significance cautiously throughout.

A.2 Inference on treatment effects based on clustering and bootstrapping

This Table considers the robustness of the main results of adding the subsidy with cafeteria-level clustering and different bootstrapping approaches.

Table A.1: Main Results Adding 5% Subsidy. Alternative Clustering Approaches

	(1)	(2)	(3)
Adding 5% Discount $\times$ Treat	0.029*** (0.001)	0.029*** (0.001)	0.029*** (0.001)
Dinner		0.017 (0.024)	0.017 (0.024)
Constant	0.357*** (0.000)	0.348*** (0.012)	0.348*** (0.012)
Cafeteria $\times$ Day-of-Week FEs	yes	yes	yes
Student FEs	yes	yes	yes
Week FEs	yes	yes	yes
Cafeteria FEs	yes	yes	yes
Clusters (standard error)	Cafeteria Level	Cafeteria Level	Cafeteria Level
Wild Bootstrap Cluster	Cafeteria Level	Cafeteria Level	Cafeteria and Week Level
Students	990	990	990
Meals	89114	89114	89114
R-squared (within)	0.0003	0.0007	0.0007

Clustered-standard errors in parentheses. Wild bootstrap p-values in square brackets.

Result (1) and (2) use Cafeteria clustering bootstraps.

Result (3) uses Cafeteria clustering with week sub-clustering bootstraps.

Cafeteria clustering requirements may be violated with few cafeterias (N=3).

OLS difference-in-difference regression. Lunch is mealtime reference for dinner.

Sample includes Balliol (treated), LMH and Exeter (controls).

Within-R2 values exclude explanatory value from FE variables

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.3 Heterogeneity with continuous baseline preference groups

The Table below uses a continuous variable for baseline vegetarian preference instead of the categorical groups used in the heterogeneity analysis. The results remain consistent, the price effect is stronger for those with higher baseline vegetarian preferences.

Table A.2: Main Results Adding 5% Subsidy. Using continuous baseline preference

	Linear Only	Quadratic Only	Linear & Quadratic Base
Adding 5% $\times$ Treat	0.032*** (0.012)	0.029* (0.017)	0.056** (0.023)
Adding 5% $\times$ Base Pref.	-0.083*** (0.019)		-0.144*** (0.030)
Adding 5% $\times$ Treat $\times$ Base Pref.	0.111*** (0.036)		0.175*** (0.052)
Adding 5% $\times$ Base Pref. (sq)		0.034 (0.053)	0.289*** (0.088)
Adding 5% $\times$ Treat $\times$ Base Pref (sq)		0.007 (0.131)	-0.309* (0.185)
Dinner	0.017*** (0.004)	0.017*** (0.004)	0.017*** (0.004)
Constant	0.347*** (0.003)	0.347*** (0.003)	0.337*** (0.004)
Cafeteria $\times$ Day-of-Week FEs	yes	yes	yes
Student FEs	yes	yes	yes
Week FEs	yes	yes	yes
Cafeteria FEs	yes	yes	yes
Clusters	Student Level	Student Level	Student Level
Students	990	990	990
Meals	89114	89114	89114
R-squared (within)	0.0011	0.0007	0.0015

ID-level clustered-standard errors in parentheses. Lunch mealtime is reference of dinner.

Base Preference is a person's average vegetarian meal consumption in year 1, centred by college mean.

Base Preference is a continuous variable, and is squared when the variable includes (sq)

OLS Difference-in-difference regression based on first 2 years of data.

Week effects are unique to each year. Day-of-week effects are unique to each cafeteria.

Within-R2 values exclude explanatory value from FE variables

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The Table below uses a continuous variable for baseline vegetarian preference instead of the categorical groups used in the heterogeneity analysis. The results remain consistent, the price effect only appears for those with higher baseline vegetarian preferences.

Table A.3: Main Results Removing 14% Subsidy. Using Continuous baseline preference

	Linear Base	Quadratic Base	Linear & Quadratic Base
Remove 14% $\times$ Treat	-0.021 (0.014)	-0.012 (0.019)	-0.032 (0.022)
Remove 14% $\times$ Base Pref.	-0.160*** (0.028)		-0.197*** (0.036)
Remove 14% $\times$ Treat $\times$ Base Pref.	-0.088* (0.049)		-0.129** (0.064)
Remove 14% $\times$ Base Pref. (sq)		-0.132* (0.072)	0.183* (0.103)
Remove 14% $\times$ Treat $\times$ Base Pref.(sq)		-0.165 (0.131)	0.123 (0.171)
Constant	0.344*** (0.002)	0.352*** (0.004)	0.337*** (0.005)
Cafeteria $\times$ Day-of-Week FEs	yes	yes	yes
Student FEs	yes	yes	yes
Week FEs	yes	yes	yes
Cafeteria FEs	yes	yes	yes
Clusters	Student Level	Student Level	Student Level
Students	855	855	855
Meals	44224	44224	44224
R-squared (within)	0.0042	0.0008	0.0047

ID-level clustered-standard errors in parentheses.

Base Preference is a person's average vegetarian meal consumption in year 1, centred by college mean.

Base Preference is a continuous variable, and is squared when the variable includes (sq)

OLS Difference-in-difference regression based on first 2 years of data.

Sample includes College B (treated), Colleges C and D (controls). Lunch only data is used.

Week effects are unique to each year. Day-of-week effects are unique to each cafeteria.

Within-R2 values exclude explanatory value from FE variables

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.4 Heterogeneity with smaller baseline preference groups

The main heterogeneity analysis splits students into three groups based on their baseline vegetarian preference. We find that in that table the estimate for the $>70\%$ group is relatively underpowered for the effect size. The Table below re-estimates the removal effect using two smaller, more extreme groups. The groups are students with a baseline vegetarian share above 80% and those with a share below 20%. Although a trade-off between sample size and contrast, the smaller group of $>80\%$ provides a larger effect size with the Minimum Detectable Effect (MDE) for 5% at 80% power. It could be argued that this smaller group is capturing reversion towards the mean for the most vegetarian students, but this would only be one sided, with no sign of reversion towards the mean among

the least vegetarian students. The results remain consistent with the main and continuous variable heterogeneity analysis.

Table A.4: Main Results Removing 14% Subsidy. Smaller baseline preference groups (>80% vs <20% veg)

	> 80% Veg	< 20% Veg
Removing 14% Subsidy $\times$ Treat	-0.082** (0.040)	0.008 (0.018)
Constant	0.876*** (0.008)	0.117*** (0.004)
Cafeteria $\times$ Day-of-Week FEs	yes	yes
Student FEs	yes	yes
Week FEs	yes	yes
Cafeteria FEs	yes	yes
Clusters	Student Level	Student Level
Students	136	308
Meals	4865	14610
R-squared (within)	0.0030	0.0000

Clustered-standard errors in parentheses.

OLS Difference-in-difference regression based on first 2 years of data.

Sample includes College B (treated), Colleges C and D (controls). Only lunches are included.

Week effects are unique to each year. Day-of-week effects are unique to each cafeteria.

Within-R2 values exclude explanatory value from FE variables

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.5 Heterogeneity with quartile groups

The main heterogeneity analysis splits students into three groups based on their baseline vegetarian preference. The Table below re-estimates the heterogeneity regressions for the subsidy addition and removal using quartiles.

Table A.5: Heterogeneity Analysis Adding 5% Subsidy. Quartiles

	Top Quartile (most veg)	Second Quartile	Third Quartile	Bottom Quartile (least veg)
Adding 5% Discount $\times$ Treat	0.078*** (0.023)	0.055 (0.036)	0.002 (0.022)	-0.016 (0.014)
Dinner	0.003 (0.008)	0.020** (0.009)	0.031*** (0.008)	0.009 (0.006)
Constant	0.795*** (0.005)	0.414*** (0.006)	0.209*** (0.005)	0.095*** (0.004)
Cafeteria $\times$ Day-of-Week FEs	yes	yes	yes	yes
Student FEs	yes	yes	yes	yes
Week FEs	yes	yes	yes	yes
Cafeteria FEs	yes	yes	yes	yes
Clusters	Student Level	Student Level	Student Level	Student Level
Students	247	246	249	248
Meals	16745	26489	24857	21023
R-squared (within)	0.0020	0.0010	0.0012	0.0004

Clustered-standard errors in parentheses. Lunch mealtime is reference of dinner.

OLS Difference-in-difference regression based on first 2 years of data.

Sample includes College A (treated), Colleges C and D (controls).

Quartiles based on each student's baseline-year (21-22) vegetarian share.

Baseline veg-share ranges: Top 1–0.6, Second 0.595–0.321, Third 0.320–0.151, Bottom 0.151–0.

Week effects are unique to each year. Day-of-week effects are unique to each cafeteria.

Within-R2 values exclude explanatory value from FE variables

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.6: Heterogeneity Analysis Removing 14% Subsidy. Quartiles

	Top Quartile (most veg)	Second Quartile	Third Quartile	Bottom Quartile (least veg)
Removing 14% Subsidy $\times$ Treat	-0.038 (0.037)	-0.090*** (0.028)	-0.016 (0.023)	-0.005 (0.023)
Constant	0.769*** (0.006)	0.406*** (0.004)	0.202*** (0.003)	0.095*** (0.005)
Cafeteria $\times$ Day-of-Week FEs	yes	yes	yes	yes
Student FEs	yes	yes	yes	yes
Week FEs	yes	yes	yes	yes
Cafeteria FEs	yes	yes	yes	yes
Clusters	Student Level	Student Level	Student Level	Student Level
Students	213	209	219	214
Meals	8006	15675	11533	9010
R-squared (within)	0.0004	0.0018	0.0001	0.0000

Clustered-standard errors in parentheses. Only lunches are included.

OLS Difference-in-difference regression based on first 2 years of data.

Sample includes College B (treated), Colleges C and D (controls).

Quartiles based on each student's baseline-year (21-22) vegetarian share.

Baseline veg-share ranges: Top 1–0.577, Second 0.571–0.287, Third 0.286–0.119, Bottom 0.119–0.

Week effects are unique to each year. Day-of-week effects are unique to each cafeteria.

Within-R2 values exclude explanatory value from FE variables

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

B Appendix: Results: Shorter Period

A potential concern is compositional change in cafeteria users during the final weeks of Spring Term, when exam schedules and end-of-year social events may alter who is eating in college. To address this, we re-estimate the main results and heterogeneity analysis using a shorter dataset that excludes the last three weeks of Spring Term in each academic year.

B.1 ATE Analysis with shorter spring terms

Table A.7: Main Results Adding 5% Subsidy. Shorter Spring Terms

	(1)	(2)	(3)
Adding 5% Discount $\times$ Treat	0.035*** (0.012)	0.035*** (0.012)	0.035*** (0.000)
Dinner		0.019*** (0.004)	0.019 (0.026)
Constant	0.356*** (0.002)	0.347*** (0.003)	0.347*** (0.013)
Cafeteria $\times$ Day-of-Week FEs	yes	yes	yes
Student FEs	yes	yes	yes
Week FEs	yes	yes	yes
Cafeteria FEs	yes	yes	yes
Clusters	Student Level	Student Level	Cafeteria Level
Students	970	970	970
Meals	80040	80040	80040
R-squared (within)	0.0004	0.0008	0.0008

Clustered-standard errors in parentheses. Lunch mealtime is reference of dinner.

Cafeteria Clustered-standard errors are not bootstrapped in this table.

OLS Difference-in-difference regression based on first 2 years of data.

Spring terms shortened by 3 weeks in both years to exclude exam period.

Week effects are unique to each year. Day-of-week effects are unique to each cafeteria.

Within-R2 values exclude explanatory value from FE variables.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.8: Main Results Removing 14% Subsidy. Shorter Spring Terms

	(1)	(2)
Removing 14% Discount $\times$ Treat	-0.037** (0.016)	-0.037* (0.012)
Constant	0.351*** (0.003)	0.351*** (0.002)
Cafeteria $\times$ Day-of-Week FEs	yes	yes
Student FEs	yes	yes
Week FEs	yes	yes
Cafeteria FEs	yes	yes
Clusters	Student Level	Cafeteria Level
Students	826	826
Meals	39287	39287
R-squared (within)	0.0004	0.0004

Clustered-standard errors in parentheses.

Cafeteria Clustered-standard errors are not bootstrapped in this table.

OLS Difference-in-difference regression based on first 2 years of data.

Spring terms shortened by 3 weeks in both years to exclude exam period.

Sample includes College B (treated), Colleges C and D (controls). Lunch only data is used.

Week effects are unique to each year. Day-of-week effects are unique to each cafeteria.

Within-R2 values exclude explanatory value from FE variables

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

B.2 Heterogeneity Analysis by baseline preference with shorter spring terms

Table A.9: Heterogeneity Analysis Adding 5% Subsidy. Shorter Spring Terms

	>70% Veg	30-70% Veg	<30% Veg
Adding 5% Discount $\times$ Treat	0.065*** (0.024)	0.056* (0.030)	0.006 (0.014)
Dinner	0.004 (0.008)	0.019** (0.009)	0.023*** (0.005)
Constant	0.838*** (0.006)	0.429*** (0.006)	0.145*** (0.003)
Cafeteria $\times$ Day-of-Week FEs	yes	yes	yes
Student FEs	yes	yes	yes
Week FEs	yes	yes	yes
Cafeteria FEs	yes	yes	yes
Clusters	Student Level	Student Level	Student Level
Students	192	313	465
Meals	12196	28852	38992
R-squared (within)	0.0017	0.0010	0.0009

Clustered-standard errors in parentheses. Lunch mealtime is reference of dinner.

OLS Difference-in-difference regression based on first 2 years of data.

Spring terms shortened by 3 weeks in both years to exclude exam period.

Week effects are unique to each year. Day-of-week effects are unique to each cafeteria.

Within-R² values exclude explanatory value from FE variables

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.10: Heterogeneity Analysis Removing 14% Subsidy. Shorter Spring Terms

	>70% Veg	30-70% Veg	<30% Veg
Removing 14% Subsidy $\times$ Treat	-0.052 (0.042)	-0.092*** (0.028)	-0.002 (0.017)
Constant	0.840*** (0.007)	0.442*** (0.004)	0.157*** (0.003)
Cafeteria $\times$ Day-of-Week FEs	yes	yes	yes
Student FEs	yes	yes	yes
Week FEs	yes	yes	yes
Cafeteria FEs	yes	yes	yes
Clusters	Student Level	Student Level	Student Level
Students	150	261	415
Meals	5146	15863	18278
R-squared (within)	0.0010	0.0018	0.0000

Clustered-standard errors in parentheses.

OLS Difference-in-difference regression based on first 2 years of data.

Spring terms shortened by 3 weeks in both years to exclude exam period.

Week effects are unique to each year. Day-of-week effects are unique to each cafeteria.

Sample includes College B (treated), Colleges C and D (controls). Lunch only data is used.

Within-R2 values exclude explanatory value from FE variables

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

C Appendix: Pre-trend of vegetarian meals

C.1 Parallel trends assessment

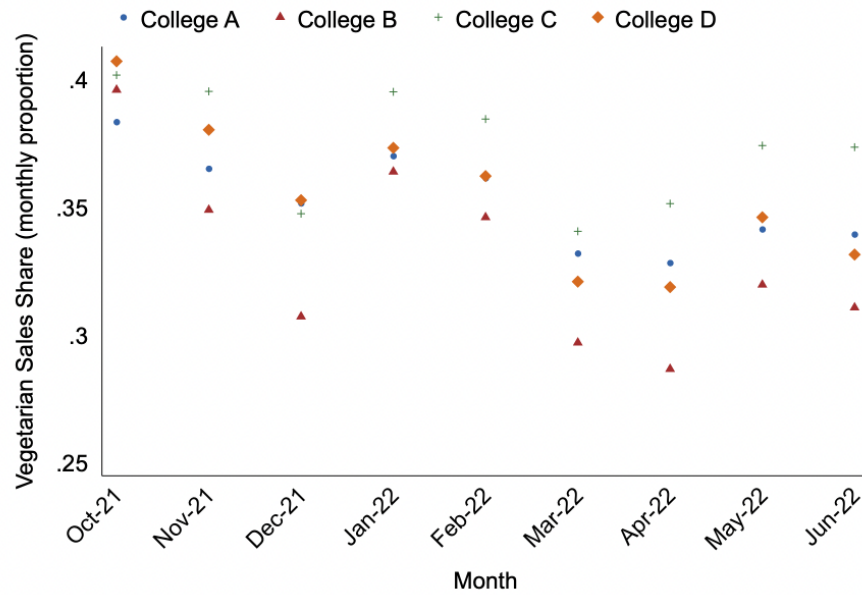


Figure A.1: Monthly Vegetarian Uptake in Baseline Year 21-22

Notes. Points plot adjusted monthly probabilities of selecting a vegetarian/vegan main by college.

Mean proportions are “adjusted” from OLS model with individual fixed effects and the controls from equation (1). ID-level clustered standard errors are used in estimation.

C.2 Decomposing the Decline in Vegetarian Demand: Within-Student versus Cohort Composition

Because each academic year brings new entrants and graduates, aggregate trends in vegetarian meal uptake could reflect within-student preference change, shifts in the dietary composition of incoming cohorts, or both. Understanding which margin drives the observed decline matters for interpreting our difference-in-differences estimates, which rely on within-student variation in the fixed-effects specification. We therefore decompose the trend using data from the two control cafeterias, where no pricing policy changes occur during the study period.

Figure A.2 plots aggregate vegetarian uptake (as a share of total sales) in control Colleges C and D (both continuously subsidised). Both exhibit a secular downward trend over the study period, irrespective of pricing regime. The same downward trend is present in the treatment colleges. In the difference-in-differences specification, common year-on-year shifts are captured by the time fixed effects; the identifying assumption is that any remaining differential trend between treatment and control colleges is attributable to the intervention, which the pre-trend evidence in Appendix C supports.

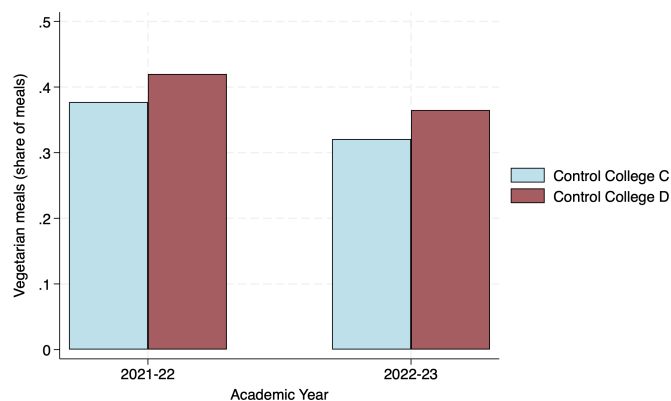
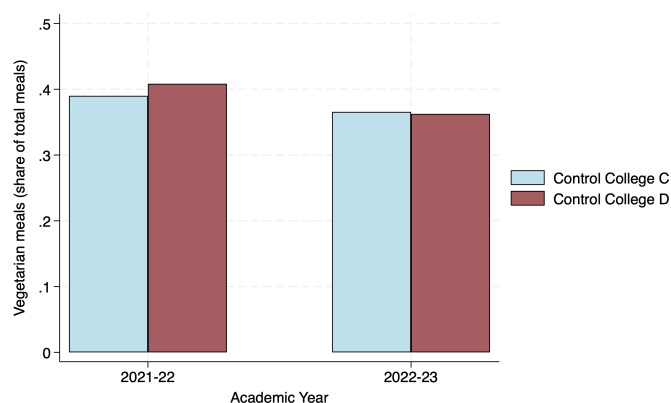


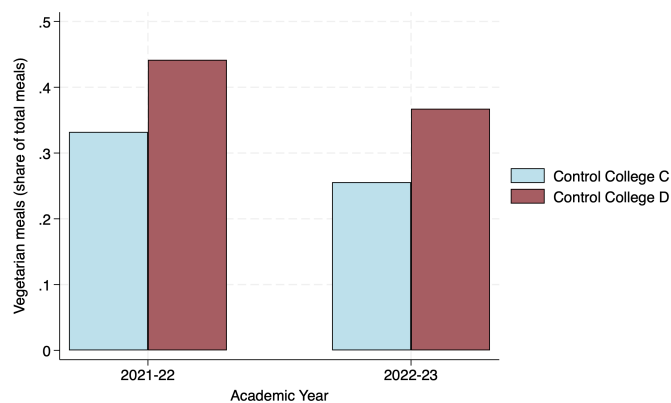
Figure A.2: Aggregate vegetarian meal share in control colleges, by academic year

To disentangle within-student preference change from cohort-composition effects,

we split the sample into (a) the balanced panel of students observed in both academic years—the same population that identifies our fixed-effects estimates—and (b) first-year entrants (“freshers”) observed in their entry year only. Panel (a) of Figure A.3 shows that even within the balanced panel, the vegetarian share declines from Year 1 to Year 2 in both control colleges, indicating genuine within-student preference drift. Panel (b) reveals that incoming cohorts also enter with progressively lower vegetarian shares. Together, the two panels show that the aggregate decline is not an artefact of either margin alone.



(a) Balanced panel: students observed in both years (within-student margin)



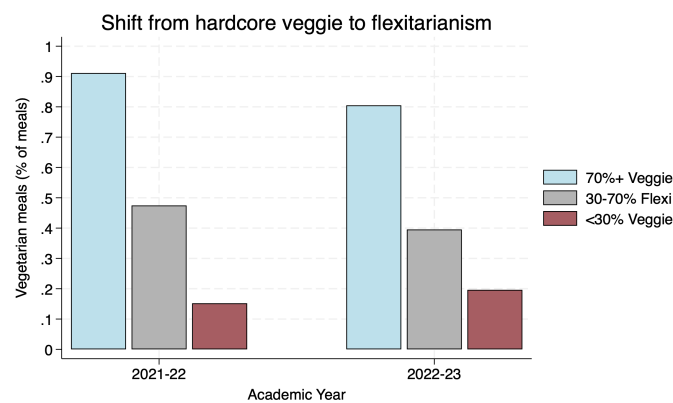
(b) First-year entrants only (cohort-composition margin)

Figure A.3: Decomposing the decline in vegetarian uptake: within-student versus cohort composition

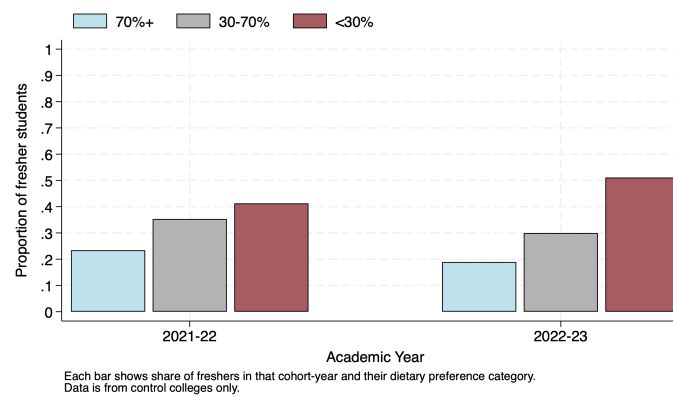
We further decompose the trend by baseline dietary preference, grouping students

by their vegetarian lunch share in their first observed year: most vegetarian ($\geq 70\%$), flexitarian (30–70%), and least vegetarian ($< 30\%$). Among the balanced-panel students in control colleges (Panel (a) of Figure A.4), the most vegetarian and flexitarian groups show a declining vegetarian share across the study period, while the least vegetarian group is largely stable. This within-student drift is concentrated among higher-baseline students—the same group that drives the heterogeneous treatment effects documented in Section 4.

Among freshers (Panel (b) of Figure A.4), successive cohorts enter with fewer students in the most-vegetarian and flexitarian categories. The cohort-composition margin therefore reinforces the within-student decline, but operates through a different channel: entry preferences, rather than preference evolution.



(a) Balanced panel: within-student transitions across dietary groups



(b) First-year entrants: cohort-level shifts in dietary composition at entry

Figure A.4: Decline in vegetarian demand by baseline dietary group: within-student drift versus cohort-composition shifts (control colleges)

D General Information About Meals and Colleges

D.1 Outcome variable, by year and college

Table A.11: Vegetarian share and total meals by college and academic year

	Vegetarian Share	Total Meals
College A		
Year		
21-22	.321	29,022
22-23	.339	30,172
Total	.330	59,194
College B		
Year		
21-22	.321	12,638
22-23	.271	29,588
Total	.286	42,226
College C		
Year		
21-22	.451	20,030
22-23	.398	21,807
Total	.425	41,837
College D		
Year		
21-22	.379	24,377
22-23	.282	24,127
Total	.332	48,504

D.2 College Meal Prices

Table A.12: Meal Prices by college and academic year

College	Mealtime	21–22	22–23
A	1-course Lunch - Vegetarian	£2.85	£3.20
	1-course Lunch - Non-Vegetarian	£2.85	£3.35
	1-course Dinner - Vegetarian	£3.45	£3.85
	1-course Dinner - Non-Vegetarian	£3.45	£4.05
B	1-course Lunch - Vegetarian	£3.00	£3.50
	1-course Lunch - Fish	£4.00	£4.00
	1-course Lunch - Meat	£3.50	£3.50
	1-course Dinner - Vegetarian	£3.50	£3.70
	1-course Dinner - Fish	£4.00	£4.00
	1-course Dinner - Meat	£3.50	£3.70
C	1-course Lunch - Vegetarian	£3.30	£3.70
	1-course Lunch - Meat	£3.50	£3.90
	1-course Dinner - Vegetarian	£3.30	£3.70
	1-course Dinner - Non-Vegetarian	£3.50	£3.90
D	1-course Lunch - Vegetarian	£1.65	£1.80
	1-course Lunch - Non-Vegetarian	£2.30	£2.50
	1-course Dinner - Vegetarian	£2.90	£3.15
	1-course Dinner - Non-Vegetarian	£3.45	£3.75

Notes. ‘Non-vegetarian’ meal prices are for fish and meat meals where colleges do not price the two differently. College B has different pricing for meat and fish, as described in the rows above.

Chapter 5 | Paper 3

Identity Activation and Choice Architecture as Complements: Experimental Evidence

Abstract: Belief-based identity models predict that making salient the link between behaviour and self-image motivates identity-consistent action. We test whether this mechanism can resolve a fundamental tension in policy design: choice architecture interventions are effective but imposing them is contested, and voluntary uptake is low. In a pre-registered framed field experiment, we show that an identity cue raises willingness to pay for a restructured choice environment by £0.45, shifting a disproportionately large share of participants across the opt-in margin. The restructured environment reduces emissions by 26%. The identity cue shifts the decision to adopt the architecture but not the choices made within it, yet this composition channel produces meaningful aggregate gains in emissions reduction relative to deploying the choice architecture intervention alone.

Declaration: I declare, as per University of Oxford School of Geography and the Environment regulations, that I have contributed more than the required 50% to the intellectual content of this paper, and that I am the primary author.

Authors: Adam Ferris, Paul Lohmann, and Michael Clark.

The paper is included in the next pages.

Identity Activation and Choice Architecture as Complements: Experimental Evidence

Adam Ferris, Paul Lohmann, Michael Clark

Abstract

Belief-based identity models predict that making salient the link between behaviour and self-image motivates identity-consistent action. We test whether this mechanism can resolve a fundamental tension in policy design: choice architecture interventions are effective but imposing them is contested, and voluntary uptake is low. In a pre-registered framed field experiment, we show that an identity cue raises willingness to pay for a restructured choice environment by £0.45, shifting a disproportionately large share of participants across the opt-in margin. The restructured environment reduces emissions by 26%. The identity cue shifts the decision to adopt the architecture but not the choices made within it, yet this composition channel produces meaningful aggregate gains in emissions reduction relative to deploying the choice architecture intervention alone.

Keywords: Identity Beliefs, Identity Intervention, Choice Architecture, Complementarity, Framed Field Experiment, Online Grocery Shopping

1 Introduction

Identity, or self-image, is a fundamental driver of human decision-making (Akerlof and Kranton, 2000; Bénabou and Tirole, 2006, 2011). Belief-based models predict that this relationship runs in both directions: acting consistently with a self-image strengthens it, and a stronger self-image motivates further consistent action (Bénabou and Tirole, 2011; Bénabou, 2015). A central implication is that a one-off intervention activating a latent identity could, in principle, set in motion a self-sustaining cycle of behavioural change (Tonke, 2025). Empirically, identity cues have been shown to shift behaviour across a range of domains (Kessler and Milkman, 2018; Tonke, 2025; Oh, 2023; Benjamin et al., 2010; Bott et al., 2020; Cohn et al., 2014). Yet measured effects are often modest and short-lived (Allcott and Rogers, 2014; Bott et al., 2020; Cohn and Maréchal, 2016), and the conditions under which an activated identity actually translates into behaviour remain insufficiently understood (Charness and Chen, 2020).

Here, we argue that a key constraint is the choice environment itself. Every choice environment has a structure, and even when not deliberately designed, that structure exerts a powerful influence on behaviour. What options are encountered, in what order, and at what cognitive cost shapes decisions substantially and independently of underlying preferences (Simon, 1955; Johnson et al., 2012). An identity cue operates through motivated self-signalling and is most effective when the decision in question is salient and experienced as diagnostic of the self (Bénabou and Tirole, 2016; Bénabou, 2015; Tonke, 2025). Where the structure of the choice environment dominates instead, this signal is crowded out, which explains why identity cues may shift self-reported beliefs without producing commensurate behavioural change. Because decisions are shaped so substantially by the structure of the choice environment, choice architecture interventions that alter behaviour without restricting options or changing economic incentives are among the most effec-

tive behavioural tools available, with documented effects across consumer choice, financial decisions, and public health (Thaler and Sunstein, 2009; Johnson et al., 2012; Ursu, 2018). Yet choice architecture interventions face their own obstacles. Unilaterally redesigning someone’s choice environment is often contested on legitimacy grounds, and when individuals must actively opt into an alternative environment, inertia and status quo bias mean uptake is typically low (Sunstein, 2016).

We propose that identity cues and choice architecture interventions are natural complements. Rather than imposing a new environment on everyone or relying on individuals to seek one out, identity cues can activate an internal motivation to willingly adopt a new choice architecture (Bénabou and Tirole, 2016; Bénabou, 2015; Tonke, 2025), rather than attempting to change choices directly within a complex choice environment. Once adopted, the choice architecture drives behavioural change through the structural factors that would otherwise crowd out the identity cue’s direct effect. We test this complementarity by asking three questions. Does an identity cue increase willingness to opt into a choice architecture intervention? Does the choice architecture intervention change behaviour for those who adopt it? And does the identity cue affect behaviour directly, or only through its effect on uptake?

We answer these questions using a pre-registered framed field experiment with 1,087 UK participants who shop on an incentivised online supermarket platform over two consecutive weeks. The platform is stocked with roughly 9,500 real Tesco products at current prices and choices are incentivised.¹ In Week 1, all participants shop under the standard product layout, providing an incentive-compatible baseline. In Week 2, we randomise exposure to an identity cue linking each partici-

¹Grocery shopping data has previously been used to study behaviour and identities, however, these were supermarket store rather than online purchases and in relation to aggregate identities, cultural, ethnic or religious distances, and impacts from international conflicts and the Brexit referendum (Bertrand and Kamenica, 2023; Atkin et al., 2021; Pandya and Venkatesan, 2016; Nardotto and Sequeira, 2021). Online purchases are more likely to not be affected by public signalling of identities and only self-signalling.

part's prior purchases to a pro-environmental self-image, and assignment to either a standard or an eco-sorted product layout, in which items are ranked by their carbon footprint. Assignment to the layout is determined by a randomly drawn row of an incentivised multiple price list (MPL), which simultaneously elicits each participant's willingness to pay (WTP) for the eco-sorted layout. One row is randomly selected for implementation, introducing quasi-random variation in layout assignment conditional on WTP and enabling causal identification of the layout effect via inverse probability weighting. This design delivers two incentive-compatible outcomes and allows us to separately identify the effect of the cue on adoption and the effect of the layout on behaviour.

Three findings emerge. First, the identity cue increases WTP for the eco-sorted layout by approximately £0.45. This shift is modest in absolute terms, but the demand curve for the eco-sorted layout is highly elastic around zero so that a small shift in WTP moves a disproportionately large share of consumers across the opt-in margin. Second, the eco-sorted layout reduces basket CO₂^e emissions by around 26%, an effect concentrated among participants with lower search effort and those indifferent between layouts, consistent with a search-cost mechanism. Third, the identity cue has no direct effect on basket emissions: it shifts the deliberative choice of environment, but not the automatic choice of items within it. The cue's contribution to aggregate emissions operates entirely through adoption: by shifting participants along the WTP distribution toward and beyond the opt-in threshold, it enlarges the population over whom the choice architecture intervention's effect is realised. This composition channel produces a meaningful gain in aggregate emissions reduction relative to deploying the choice architecture intervention alone. We also find that the cue raises self-reported green identity, consistent with the belief-updating mechanism predicted by belief-based identity models.

This paper makes three contributions. First, we provide micro-level experimen-

tal evidence on the conditions under which the identity-behaviour feedback loop closes. Our results show that the salience and self-signal value of the decision is a key moderator. Identity cues close the loop for decisions experienced as meaningful commitments about the self, but not for choices shaped primarily by the structure of the choice environment. This reconciles the strong predictions of belief-based identity models (Bénabou and Tirole, 2006, 2011) with the modest direct behavioural effects documented empirically (Allcott and Rogers, 2014; Bott et al., 2020; Cohn and Maréchal, 2016). It also extends the work of Tonke (2025) and tracks the full belief to behaviour and back to belief chain within the same individuals: our panel design elicits self-reported identity both before and after the cue, so the within-person difference-in-differences directly identifies the belief updating that these models posit. Second, we provide, to our knowledge, the first revealed-preference estimates of demand for a choice architecture intervention, using an incentivised multiple price list to elicit willingness to pay for an alternative product ordering. We further show that a low-cost identity cue can shift a meaningful share of consumers across the opt-in margin, increasing the reach of an effective choice architecture intervention without imposing it. This contributes to the literatures on behavioural interventions and libertarian paternalism (Johnson et al., 2012; Thaler et al., 2013; Sunstein, 2016). Third, we document a general design principle for managers and policymakers: pairing a low-cost preference-activating intervention with a structural one can achieve broader coverage and more sustained impact while preserving consumer autonomy. The principle applies across any setting where a choice architecture intervention is effective but faces adoption friction, from platform design and default settings to workplace nudges and public policy.

The rest of the paper proceeds as follows. Section 2 describes the experimental design, interventions and procedure. Section 3 describes the data and sample. Section 5 sets out the estimation strategy and Section 6 reports the results. Section 7 discusses the policy implications and Section 8 concludes.

2 Experiment

In the following, we describe the experimental setting and shopping task (Section 2.1), the two interventions (Section 2.2), the experimental design (Section 2.3), and the experimental procedure (Section 2.4). The experiment was pre-registered on the AEA RCT Registry (April 2025, Clark, Ferris, Lohmann).²

2.1 Experimental Setting and Shopping Task

The experiment was conducted on Woods, an experimental online supermarket platform that replicates the interface and choice environment of a standard UK online grocery retailer. The platform draws on a dataset used in prior research (Clark et al., 2022; Pechey et al., 2022c,a) and offers roughly 9,500 real food items at real prices from Tesco, the UK’s largest grocery retailer.³ Product pictures, prices, and a combined environmental and nutritional label derived from Clark et al. (2022) were displayed on every item in both weeks, unconditional on treatment assignment. Participants therefore had access to full environmental and nutritional information throughout the study and any difference in basket composition between conditions reflects changes in the decision process rather than in the information available. Screenshots of the platform are provided in Appendix D (Figures A.8 and A.9). Further details on the label scoring system are in Appendix C.⁴

In both weeks, participants were asked to complete their typical weekly grocery

²Pre-registration: <https://doi.org/10.1257/rct.15869-1.2>. Ethics: MS IDREC Committee, reference R65010.

³Tesco holds a 28.5% share of the UK grocery market and a 35.7% share of online grocery, the largest of any UK retailer (Tesco PLC Preliminary Results 2025/26). Participants purchased 3,549 distinct food items across the two weeks of the experiment.

⁴Online grocery platforms have been used in prior economics research to study food policy interventions (Panzone et al., 2021a,b; Demarque et al., 2015; Kanay et al., 2021). Woods not only offered UK-based products, notably those available in the Tesco supermarket store (different to Demarque et al. (2015); Kanay et al. (2021) which uses France’s Casino supermarket items), it also provides our framed field experiment with the most complete grocery list of all these papers with 9,500 items. The Woods platform also uniquely included eco-score and nutri-score information derived from Clark et al. (2022).

shop on Woods. To reduce between-subject heterogeneity while preserving meaningful choice variation, they were additionally asked to include six prescribed item categories: soup, breakfast cereal, a ready meal to heat, a sandwich, a pizza, and a fruit or vegetable item.⁵ The prescribed list was visible throughout the shopping task (Figure A.8), and participants were required to spend between £25 and £50, enforced by the platform.

Choices were incentivised through a random-incentive mechanism: five participants, selected at random after data collection, received one of their two shopping baskets for real-world collection from a Tesco of their choice, together with any unspent portion of the £50 budget. Participants were informed of this lottery before beginning the shopping task. This payoff structure makes choices consequential and reduces experimenter-demand concerns (List, 2011; List and Price, 2016; Zizzo, 2010). Because the study does not focus on risky decision-making, incentivising only a subset of participants does not raise concerns about income effects on risk attitudes (Anderson et al., 2023).

The platform records all participant actions on the system (groceries added/removed) and the final basket of groceries at checkout in each wave. Our primary basket outcome is the environmental impact of the final purchased basket, measured as kilograms of CO₂-equivalent emissions (kg CO₂e), constructed by summing item-level emissions using the product-level data of Clark et al. (2022) and following Panzone et al. (2021b,a).

2.2 Interventions

The experiment studies two interventions that are designed to address different constraints. The identity cue targets the motivated reasoning stage of decision-making, making salient the link between past behaviour and a valued self-image.

⁵The prescribed categories are common in prior studies and feature substantial within-category variation in environmental impact, which improves comparability across participants.

The choice architecture intervention targets the automatic stage, reshaping the environment within which grocery decisions are made. Together, they are candidates for complementarity precisely because each operates where the other cannot.

2.2.1 Identity cue

The identity cue was shown to participants in Week 2 as a full-screen pop-up message, displayed in bright colours to attract attention. It required participants to click “Next” to proceed; no other content appeared on that screen. The message read: “Last time, you purchased eco items — you’re one of our eco-friendly shoppers! Great, keep going and we will keep providing green options.” The cue was designed around three principles. First, it referenced the participant’s own past behaviour, drawing on belief-based models in which making the identity–behaviour link salient motivates identity-consistent action (Akerlof and Kranton, 2000; Bénabou and Tirole, 2006, 2011; Bordalo et al., 2020; Tonke, 2025).⁶ The identity claim was necessarily true for all participants by design. The prescribed shopping list required inclusion of at least one fruit or vegetable item, which receives the top eco-score (A) on the platform’s six-level scale, ensuring that every participant had purchased at least one eco-friendly item in Week 1.⁷ Second, it applied a noun-based attribution — “you are one of our eco-friendly shoppers” — since noun-based identifications have been shown to generate stronger behavioural commitment than verb-based ones (Bryan et al., 2011). Third, the closing clause — “keep going and we will keep providing green options” — was included to dampen psychological reactance, since behavioural interventions can be perceived as threatening individual autonomy and may backfire if they feel manipulative (Sunstein, 2017; Osman et al., 2020; Sparkman et al., 2021).

⁶These models predict that deviations from one’s perceived identity incur a psychological cost — a negative self-signal — which creates an incentive to act consistently with the activated self-image (Bénabou and Tirole, 2006, 2011).

⁷The cue is therefore less individually tailored than one constructed from each participant’s full Week 1 basket, since the prescribed items alone were sufficient to make the message truthful. This design choice was made deliberately to avoid deception. Participants were not informed of this.

The final wording was selected through a pre-test with 50 UK participants recruited via Prolific, who rated six candidate statements on personal relevance, emotional resonance, motivational impact, clarity, and autonomy preservation. The two highest-rated statements were combined to form the cue used in the experiment. A screenshot is provided in Appendix D, Figure A.7.

2.2.2 Eco-sorting layout

The choice architecture intervention varies the order in which grocery items are presented on the platform. Under the standard layout — experienced by all participants in Week 1 — items returned by search or category browsing appear in the platform’s default order. Under the eco-sorting layout, items are ranked by their eco-score (Clark et al., 2022), placing the most environmentally friendly products first.

The sorting order is a well-established choice architecture intervention with large documented effects on consumer choice across a range of settings, from political ballot ordering to online retail (Johnson et al., 2012; Ursu, 2018; Thaler and Sunstein, 2009; Kamenica, 2012; Derakhshan et al., 2022; Lohmann et al., 2024). It operates through bounded rationality and position-dependent search costs: in online environments, the number of options far exceeds consumers’ attention and cognitive resources, so only products placed in the first few positions are readily accessible (Derakhshan et al., 2022; Simon, 1955). Consumers therefore tend to satisfice rather than optimise (Simon, 1955; Augenblick and Nicholson, 2016), making primacy a meaningful determinant of choice.

2.3 Experimental Design

The experiment runs over two consecutive weeks. In Week 1, all participants complete an incentive-compatible baseline grocery shop under the standard layout,

providing pre-treatment revealed-preference outcomes against which Week 2 outcomes are compared across all arms. In Week 2, we introduce two sources of random variation: exposure to the identity cue, randomised across three arms, and assignment to the eco-sorting or standard layout, determined quasi-randomly via an incentivised multiple price list (MPL). Participants complete a second grocery shop under the assigned layout after the MPL. This design delivers two incentive-compatible outcomes: a cross-sectional WTP measure elicited in Week 2 and a panel outcome from revealed grocery baskets observed in both weeks.

2.3.1 Treatment arms

In Week 2, participants were randomly assigned to one of three arms, with quotas ensuring an even split by gender and university education:

- *No Cue*: no identity cue is provided.
- *Pre-MPL Cue*: the identity cue is shown before the WTP elicitation.
- *Post-MPL Cue*: the identity cue is shown after the WTP elicitation but before the Week 2 shopping task.

The text and visual design of the cue was identical across the two cue arms; only its timing relative to the MPL differed. This timing variation serves a distinct identification purpose: the Pre-MPL Cue arm allows us to identify the cue's effect on WTP, while the Post-MPL Cue arm, having seen neither the cue nor the layout at the point of elicitation, provides a clean counterfactual for that comparison. Any effect of the Post-MPL Cue on basket emissions must therefore operate through a direct identity-to-shopping channel, independently of WTP.

In addition to its potential effect on our WTP outcome, the cue's internalisation is assessed through a composite self-reported green-identity measure collected in

both weeks for all arms.⁸

2.3.2 Willingness-to-pay elicitation and layout assignment

In Week 2, we elicited each participant's willingness-to-pay to have grocery search results presented under an eco-sorted layout (items ordered by environmental score) rather than the platform's default ordering, using an incentivized multiple price list (MPL; Figure A.11). Each MPL row presents a choice between the eco-sorted layout and the default layout, coupled with a monetary trade-off framed as a reduction from a £7 endowment, payable to winners of the basket lottery. Participants completed 19 of these binary choices spanning a pre-specified grid of prices ($\{7; 5; 4; 3; 2.50; 2; 1.50; 1; 0.50; 0\}$ in GBP), with the alternative fixed at £0.⁹

The MPL is incentive compatible on two margins. After participants completed the MPL, one row was randomly selected for implementation. The participant then performed the Week 2 shop under the layout associated with their choice in the selected row, and, conditional on winning the basket lottery, the corresponding price was deducted from their £7 endowment.¹⁰ Before the incentivised MPL, participants were given generic instructions on how to use an MPL (without reference to the layout choice) and were required to answer a comprehension question correctly before proceeding (Figure A.10). For example, if row 6 of Figure A.11 ("Choose option B, and keep £2.00") were drawn for a participant who had selected 'Option B: Eco Ranking', that participant would complete the Week 2 shop under the eco-sorted layout and, if a basket winner, receive £2 of the £7 endowment.

We did not restrict participants to single-switching behaviour in the MPL. In Sec-

⁸The green-identity measures do not allow us to disentangle the cue's operation as an identity prompt from related channels such as moral reminders or pro-environmental priming. This does not affect the interpretation of our central result.

⁹We use non-negative, monotonically ordered prices within each sorting-layout option to reduce participant confusion (Andersen et al., 2006; Epperson and Gerster, 2026). The price range was chosen to exceed typical WTP values observed in our control group (see Figure A.1).

¹⁰This payment rule makes MPL choices consequential for the shopping environment (layout) and, conditional on winning, for monetary payoffs, discouraging inflated WTP reports.

tion 3 we describe how we treat the small number of inconsistent response patterns (e.g., multiple switching) in the construction of our WTP measures.

The row-draw probabilities were not uniform. The two corner rows — keeping the full £7 under either layout, implying a £0 WTP for the alternative — were each drawn with probability $18/53 \approx 34\%$, while each of the seventeen interior rows was drawn with probability $1/53 \approx 2\%$, making the corner rows 18 times more likely to be drawn than any interior row.¹¹ As a result, 68% of implemented draws came from these two corner rows. Because almost all participants are willing to accept either layout when it is free, the realised Week 2 layout assignment is, for the large majority of participants, independent of their true WTP, removing endogenous self-selection into the sorting layout. The identification strategy that exploits this quasi-randomisation is described in Section 5.

2.4 Procedure

The study was conducted online using Prolific for recruitment and payments, Qualtrics for surveys and treatment assignment, and Woods for the shopping tasks. The study flow is illustrated in Figure 1.

2.4.1 Recruitment and attrition management

Participants were recruited via Prolific and informed upfront that the study required two sessions over consecutive weeks, with the dates of both sessions communicated in advance. Before proceeding, participants were asked to confirm that they expected to be available for both weeks, serving as a commitment device. Each completed session was compensated at £2 for an estimated 18-minute task. To further reduce attrition, participants returning for Week 2 received an additional

¹¹Keep £7.00 for Option B” had a 34% chance of being drawn, while keep £1.00 for Option B” had a 2% chance. Participants with abnormally high WTP to avoid a given layout would always select the corner row regardless of the draw. These are excluded from the basket-emissions analysis as described in Section 3.

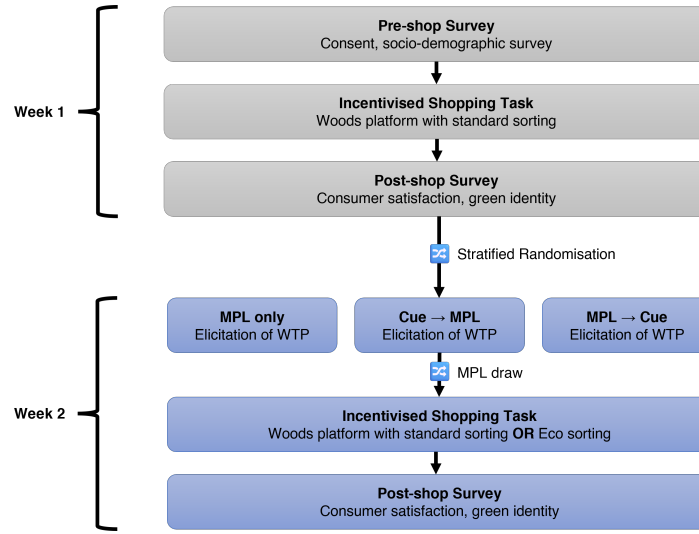


Figure 1: Laboratory Procedure for Participants

Note: This figure illustrates the order of the different sections of the study.

£1 bonus, and reminder messages were sent via Prolific ahead of the second session. All participants were entered into the basket lottery described in Section 2.1. In Week 2, participants were additionally informed that lottery winners would receive a £7 endowment, the realised value of which would depend on their MPL responses.

2.4.2 Week 1

Participants were directed from Prolific to a Qualtrics survey where they received study details and provided informed consent. They then completed a survey covering socio-demographic characteristics (age, gender, education, household size, income, and typical weekly grocery budget). Following the survey, participants were directed to Woods, where they received task instructions before completing the shopping task (see Appendix D.4). After checkout, they were redirected to a post-shop Qualtrics survey collecting shopping experience items, the green-identity measure, and filler questions on health and exercise included to reduce demand effects (Zizzo, 2010). Contextual measures (hunger, shopping location) and instru-

ments for locus of control, self-control, and the cognitive reflection test were also collected at this stage.

2.4.3 Week 2

Participants who completed Week 1 were re-invited via Prolific and provided with a link to the Week 2 survey. Because treatment arms are assigned in Week 2, attrition between weeks is independent of treatment by construction; attrition details are reported in Section 3. On entering the survey, participants were randomly allocated to one of the three treatment arms, with quotas on gender and education. After reconfirming consent, all participants received the MPL instructions and completed the comprehension test; those who did not pass on the first attempt were permitted to retake it. Participants in the Pre-MPL Cue arm saw the identity cue before the MPL; those in the Post-MPL Cue arm saw it immediately after. Following the MPL and the cue (where applicable), participants completed the Week 2 shopping task on Woods under the assigned layout, subject to the same budget and prescribed item requirements as in Week 1. After checkout, they completed a post-shop survey repeating the satisfaction, green-identity, and filler items from Week 1.

3 Data

3.1 Sample

Participants were recruited via Prolific. Eligibility filters restricted participation to United Kingdom residents aged 18 or older who reported being the primary grocery shopper in their household and having prior online grocery-shopping experience. To preserve a sufficiently broad feasible choice set on the Woods platform (roughly 9,500 products), we excluded participants reporting dietary constraints that would severely limit basket composition (vegan, halal, or kosher). Participants following other diets (e.g., vegetarian, pescatarian, or calorie-controlled) remained eligible.

Recruitment opened on 30 April 2025, and Week 2 data collection closed on 14 May 2025.

The planned Week 1 recruitment target was 1,045 participants, based on power calculations for both the cross-sectional and panel outcomes, budget constraints, and prior studies with related designs. The target incorporated an anticipated 15% non-completion/attrition margin. We recruited 1,087 participants in Week 1.¹² After excluding individuals who did not complete both weeks, did not complete mandatory survey items, or failed the WTP cleaning criteria,¹³ the final balanced two-week panel comprises 875 participants observed in both weeks.

Table 1 summarises the sample construction. The two primary outcomes are estimated on different subsets of the balanced panel, for the identification reasons set out in Section 5: the cross-sectional WTP analysis uses the 850-participant sample with at most three MPL preference reversals ($n_{\text{changes}} \leq 3$, 97% of the panel), while the panel basket-emissions analysis uses an inverse-probability-weighted (IPW) DiD sample of 778 single-switchers, obtained by dropping the 49 always-A and always-B non-switchers and the single corner switcher whose MPL responses imply $|\text{WTP}| = £7$, where the design propensity score equals 0 or 1. Appendix A re-estimates the WTP regression on the 828-participant single-switcher subsample; results are comparable.

Within the final two-week balanced panel sample, 98% (858 participants) passed the Week 1 attention check, and 93% passed the Week 2 MPL comprehension check on the first attempt. These participants are retained in the main analysis; robustness checks excluding them are reported in Appendix A.

Table 2 reports key socio-demographic characteristics, baseline shopping behaviour,

¹²Recruitment was extended slightly because Week 1 completion was faster than anticipated.

¹³Appendix E.4 details the MPL cleaning rules, including the treatment of multiple-switch and no-switch response patterns.

Table 1: Sample Construction

	<i>N</i>	% of Week 1
Week 1 recruited (Prolific)	1,087	100.0%
Balanced two-week panel ^a	875	80.5%
WTP analysis sample ($n_{\text{changes}} \leq 3$) ^b	850	78.2%
Single-switchers and non-switchers ($n_{\text{changes}} \leq 1$)	828	76.2%
IPW-DiD basket-emissions sample ^c	778	71.6%

Notes: ^aParticipants completing completing the Week 1 and Week 2 surveys, all mandatory survey items, and a valid checkout in both weekly grocery-shopping tasks. ^b Excludes participants with more than three preference reversals in the Multiple Price List (MPL); see Appendix E.4 and Table A.14. The estimated WTP regression sample is 849 after list-wise deletion of one observation with missing covariates. ^c Restricts to single-switchers ($n_{\text{changes}} = 1$) whose design-known propensity of being allocated to the eco-sorted layout lies strictly between 0 and 1: drops both the always-A and always-B non-switchers at $n_{\text{changes}} = 0$ (49 participants), and one corner single-switcher with $|\text{WTP}| = £7$; see Section 5.

and balance tests across treatment arms.¹⁴ In the analysis sample, 50% of participants are female; 23% are younger than 31 years, 32% are 32–41, 21% are 42–52, and 23% are older than 52 years. Sixty-six percent hold a university degree, 46% report annual household income below £30,000, and 20% reporting £50,000.

At baseline (Week 1), the mean basket carbon footprint is 7.29 kg CO_2e (median: 6.56 kg); the distribution is right-skewed. Participants spend on average £40 within the enforced £25–£50 budget, purchase roughly 18 items, and spend 11 minutes on the shopping task (median: 9 minutes). Additional full-sample descriptive statistics are reported in Table A.1 (Appendix A).

Consistent with random assignment, we do not detect statistically significant baseline differences across treatment arms in observed Week 1 characteristics (Tables 2 and A.2).¹⁵

¹⁴A more comprehensive balance table with additional baseline variables is reported in Table A.2, Appendix A.

¹⁵Treatment-arm quotas were imposed on gender and university-education attainment.

Table 2: Descriptive statistics and randomisation balance checks.

	Control n=295 (33.7%)	Pre-MPL Cue n=296 (33.8%)	Post-MPL Cue n=284 (32.5%)	Test
Female	146 (49.5%)	150 (50.7%)	140 (49.3%)	0.937
Age: less than 31 y.o.	61 (20.7%)	68 (23.0%)	73 (25.7%)	0.627
University Educated	200 (67.8%)	192 (64.9%)	186 (65.5%)	0.731
Income Group: less than £30k	138 (46.8%)	122 (41.2%)	145 (51.1%)	0.188
Low Green Self Wk1	150 (50.7%)	152 (51.4%)	132 (46.5%)	0.458
Basket Total GHG (kg CO2e)	7.41 (3.99)	7.31 (4.16)	7.14 (3.69)	0.704
Search Effort (mins)	10.77 (7.23)	11.93 (21.51)	11.02 (10.01)	0.585

Notes: Joint significance balance tests across all three treatment arms are reported. Categorical variables are reported with number of participants and in parentheses the share in category = 1. P-value from Pearson chi-squared test. 2 category variables: Female, University Educated, Low Green Self Wk1. 3 category variables: Income Group. 4 category variables: Age Group. For continuous variables: Mean (SD); p-value from linear regression.

3.2 Variables

3.2.1 Primary Variables

The analysis has two incentive-compatible primary outcomes: the willingness-to-pay for the eco-sorted layout (elicited in Week 2) and the environmental impact of the grocery basket, observed in both weeks.

Willingness-to-pay for eco-sorting. Willingness-to-pay (WTP) for the eco-sorted layout is the maximum premium, in GBP, that a participant is willing to pay to shop under the eco-sorted rather than the standard layout. Positive values indicate a preference for eco sorting; negative values indicate that the participant requires a subsidy to accept it (equivalently, prefers standard sorting at equal prices). WTP is elicited in Week 2 before the shopping task using the incentivised 19-row MPL described in Section 2.3, and is bounded in $[-£7, £7]$. It is constructed using the last-switch rule: the highest premium at which a participant still chooses eco sorting before switching to standard sorting. Full coding rules and robustness details are reported in Appendix E.

The main analysis retains the 850 participants with at most three preference switches in the MPL, following evidence that multiple switching need not indicate low-quality responses (Yu et al., 2021); see Table A.14. As a robustness check we also report results on the stricter single-switcher and non-switcher subsample (828 participants).

Grocery basket environmental impact. The environmental impact of a participant i 's Week- t basket, Y_{it} , is the total kilograms of CO₂-equivalent emissions (kg CO₂e) attributable to the basket chosen at checkout on the Woods platform. Higher values correspond to less environmentally sustainable baskets. We construct Y_{it} by summing item-level emissions in the final basket, using the product-level data of Clark et al. (2022). Item-level emissions are available for all products on the Woods platform, so there is no within-basket missingness. The variable is defined for the balanced two-week panel of 875 participants who completed a valid checkout in both weeks (see Section 3 for the sample-flow table).

The main analysis uses the natural log of basket emissions, $\ln(Y_{it})$, as the dependent variable, following prior grocery-emissions research (Panzone et al., 2021a,b). The log transformation accounts for the right-skewed distribution of week-level basket emissions (shown in Appendix Figure A.5), mitigates heteroskedasticity, and yields coefficients interpretable as approximate percentage changes.¹⁶

Treatment indicators. The analysis exploits two sources of exogenous variation introduced in Week 2. Cue_i equals one for participants assigned to either cue arm (pre- or post-MPL) and zero for the no-cue control. To separate the two cue arms in the cross-sectional WTP analysis, we further define $\text{Cue}_i^{\text{pre}}$ and $\text{Cue}_i^{\text{post}}$ equal to one for participants who received the cue before and after the MPL, respectively. For

¹⁶While we initially pre-registered basket emissions per 100g purchased as the primary outcome variable, we instead use total basket emissions, which better captures real-world basket-level impact, following Panzone et al. (2021b,a). Product weight data are incomplete for some items, preventing construction of a comprehensive per-100g measure for the full sample.

the WTP outcome the relevant regressor is $\text{Cue}_i^{\text{pre}}$, since Post-MPL-cue participants have not yet seen the cue at elicitation and are therefore pooled with the No-Cue arm.

EcoSorting_i equals one if the participant completed the Week 2 shop under the eco-sorted layout and zero under the standard layout (the layout faced by all participants in Week 1). Layout assignment is determined by the randomly drawn MPL row. Because the extreme MPL rows are drawn with disproportionately high probability (Section 2.3), layout assignment is independent of WTP for all but the corner consistent-switchers; we exploit this conditional randomisation as the basis for a quasi-random comparison of eco versus standard sorting, with the identification details set out in Section 5.

3.2.2 Secondary Variables

The remaining variables enter the analysis as controls in the panel emissions regressions, as conditioning variables in the heterogeneity analysis, or as secondary outcomes. Exact survey wording, response scales, and coding rules are reported in Appendix G.

Time-invariant covariates In Week 1 we collect age, gender, number of children, household income, and educational attainment, together with two shopping-context covariates, the participant’s typical weekly grocery budget and whether they frequently shop for others. Because the main specifications include individual fixed effects, these variables do not enter the panel regressions as controls; they are used in the heterogeneity analysis and the cross-sectional WTP specifications.

Green identity. Pro-environmental self-identity captures the extent to which participants view sustainable consumption as part of who they are. We elicit it in the post-shop survey in both weeks using three 11-point items adapted from the

self-identity scales of Sparks and Shepherd (1992) and Cook et al. (2002): (i) the importance of being a sustainable consumer to one's self-concept, (ii) whether the participant regards themselves as someone who cares about the environment and sustainability, and (iii) whether they see themselves as a sustainable consumer. The three responses are summed to form a green-identity score ranging from 3 to 33, with higher values indicating a stronger pro-environmental self-concept. The Week 1 score serves as a baseline moderator in the heterogeneity analysis of the WTP and basket outcomes. The Week 1-to-Week 2 change is used as a secondary outcome to assess whether the identity cue and the eco-sorted layout shift self-reported green identity.

Shopping behaviour. We record three week-varying measures of shopping behaviour on the Woods platform: the number of items in the final basket, total basket spend (GBP), and search effort, defined as minutes spent on the shopping task (excluding survey time). Items and spend enter the emissions regressions as controls. The Week 1 value of search effort is used as a type-of-consumer covariate in the heterogeneity analysis, and the Week 1-to-Week 2 change is a secondary outcome when estimating the effect of the treatments on search effort.

4 Hypotheses

Our framed field experiment is designed to identify two pre-registered causal effects: the impact of the identity cue on willingness to pay for the eco-sorted layout (H1) and the impact of the eco-sorted layout on basket emissions (H2). Each is grounded in a distinct theoretical literature and accompanied by secondary tests on mechanism and heterogeneity. One hypothesis, H2c, was not pre-specified and is flagged as exploratory throughout.

H1: Identity cue and demand for the eco-sorting layout

Identity economics models utility as partly derived from acting consistently with a self-image (Akerlof and Kranton, 2000), with deviations from one’s perceived identity incurring a psychological cost, a negative self-signal (Bénabou and Tirole, 2006, 2011; Charness and Chen, 2020). A cue that links past behaviour to a valued identity type raises the salience of that self-image, increasing the marginal value of identity-consistent actions (Tonke, 2025), and the cost of actions that now saliently diverge from it (Saccardo and Serra-Garcia, 2023). By the same logic, agents should also value tools that make identity-consistent behaviour easier to achieve (Bénabou and Tirole, 2004).

Identity updating is unlikely to be monotonic in baseline identity: the marginal value of an additional identity-confirming signal is largest for agents in intermediate states, neither already strongly aligned nor strongly misaligned with the trait being cued (Bénabou and Tirole, 2011). We return to this non-monotonicity in H1c.

In Week 2, participants in the two cue arms receive a message linking their Week 1 basket to a pro-environmental self-image (“you’re one of our eco-friendly shoppers”). Cued participants should therefore place greater value on the eco-sorted layout, which facilitates identity-consistent behaviour by lowering the search cost of greener items.¹⁷

H1a (Main Effect, pre-registered): Participants exposed to the identity cue before the MPL (Pre-MPL Cue arm) report a higher willingness to pay for the eco-sorted layout than participants in the No-Cue and Post-MPL-Cue arms.

H1b (Mechanism Check: Green-Identity Updating, pre-registered): The cue raises self-reported green identity from Week 1 to Week 2, consistent with the interpretation that it operates through motivated self-signalling.

¹⁷The cue is designed to dampen the risk of reactance; see Section 2.3.

H1c (Heterogeneity Analysis: Green Identity Baseline, pre-registered): The cue effect on WTP is largest for participants with an intermediate Week 1 green-identity score, and weaker at both the high and low ends of the distribution.

H2: Eco-sorting layout and grocery basket emissions

Consumers satisfice rather than optimise: they face time and search-effort constraints (Simon, 1955) and are therefore influenced by the choice architecture of the shopping environment rather than processing all available information (Johnson et al., 2012; Weber, 2013). Search-effort models in turn highlight how the position of a product in a sequential list shapes the probability that it is chosen, and how much rent prime positions can extract (Stigler, 1961; Ursu, 2018).

Re-sorting the product list acts on consumption choices through two complementary channels. First, by placing greener products in prime positions, the eco-sorting layout lowers the search cost of finding them, so that a satisficing consumer is more likely to add them to their basket before the search-effort budget is exhausted (Ursu, 2018; Johnson et al., 2012; Weber et al., 2007; Johnson et al., 2007). Second, the eco-sorted ordering raises the salience of the environmental dimension of the choice set (Bordalo et al., 2013, 2020), which can shift weighting toward that attribute. Our design identifies the combined effect of both channels; the heterogeneity analysis in H2b provides indirect evidence on their relative contribution.

H2a (Main Effect, pre-registered): Participants assigned to the eco-sorted layout in Week 2 choose baskets with lower CO₂e emissions than those assigned to the standard layout.

H2b (Mechanism: Search Effort, pre-registered): The layout effect on basket emissions is largest for participants with shorter Week 1 search effort, consistent with a search-cost channel in which low-effort shoppers are most exposed to what-

ever the layout places first.

H2c (Heterogeneity Analysis: WTP and Willful Ignorance, exploratory): Participants who revealed a preference to avoid the eco-sorted layout (negative WTP) exhibit larger basket-emission reductions when nonetheless assigned to it – consistent with a willful-ignorance interpretation in which agents pay for cognitive flexibility because they anticipate the layout will move their behaviour (Saccardo and Serra-Garcia, 2023; Epperson and Gerster, 2026).¹⁸

Managerial and policy evaluation of the interventions. Beyond the identification questions in H1 and H2, the identity cue and the eco-sorted layout are both candidate tools a retailer or policy-maker could deploy at scale. Because the pre-registered analysis was designed to speak to deployment as well as to identification, we also evaluate each intervention along managerial and policy-relevant dimensions: unintended effects on basket spend and shopping time, the distribution of effects across consumer types (who benefits and who is left unaffected or made worse off), and the likely persistence of the effect once the nudge is removed. This evaluation is structured around the FORGOOD ethical framework (Lades and Delaney, 2022) and the scalability checklist of Al-Ubaydli et al. (2017, 2020), and is presented in Section Section 7.

5 Estimation Strategy

To tackle our two distinct primary outcome variables, we use two different pre-registered estimation strategies. The willingness-to-pay (WTP) is measured once per participant in Week 2, so we use a cross-sectional OLS regression to estimate the effect of the identity cue on WTP for the eco-sorting layout. In contrast, the

¹⁸This hypothesis was not part of the pre-analysis plan and was motivated by the contemporaneous willful-ignorance literature; we report it as suggestive mechanism evidence rather than as a confirmatory test.

grocery basket emissions are observed in both Weeks 1 and 2 for each participant, allowing us to use a panel data approach with individual fixed effects to estimate the effect of the eco-sorted layout on within-person changes in emissions between the two weeks.

The two strategies are estimated on different subsets of the 875-participant balanced two-week panel, reflecting the different outcomes' identification requirements. The cross-sectional WTP analysis uses the 850 participants with at most three preference reversals in the MPL ($n_{\text{changes}} \leq 3$), since multiple-switching beyond this threshold is treated as evidence of MPL inattention or comprehension failure (see Section 3 and Appendix E).¹⁹ The basket-emissions panel analysis restricts the initial sample to the 828 single-switchers and non-switchers ($n_{\text{changes}} \leq 1$), and then trims the consistent-corner participants at $|\text{WTP}| = \pounds 7$ for whom the propensity score is exactly zero or one and the layout assignment is non-random.²⁰ The resulting estimation sample for the eco-sorting effect on baskets comprises 778 participants. The identification strategy and IPW specification are described in Section 5.2. Robustness checks reported in Appendix A re-estimate the WTP regression on the same 828-participant single-switcher sample to confirm that comparisons across the two outcomes are not driven by sample-construction differences.

5.1 Willingness-to-Pay Estimation Strategy

Willingness-to-pay for the eco-sorting layout is elicited once per participant in Week 2 (Section 2.3), so the cue effect on WTP is identified from a cross-sectional comparison across treatment arms. Identification rests on the random assignment of the identity cue in Week 2, with quotas (50-50) only on gender and university education to guarantee balance on these two assumed predictive covariates given our

¹⁹One observation is lost to listwise deletion of a missing covariate, so the estimated WTP regression sample is 849.

²⁰The trim drops the 49 always-A and always-B non-switchers at $n_{\text{changes}} = 0$, and the one corner single-switcher whose responses imply $|\text{WTP}| = \pounds 7$. See examples in Appendix Table A.16.

moderate sample size. Balance on the remaining Week 1 covariates is reported in Table 2.

The treatment indicator is $\text{Cue}_i^{\text{pre}}$, which equals one for participants assigned to the Pre-MPL Cue arm and zero otherwise. Pooling the No-Cue and Post-MPL Cue arms in the comparison group is justified because at the moment WTP is elicited, Post-MPL Cue participants have not yet seen the cue and are therefore no different from the No-Cue control for this outcome.

We estimate

$$\text{WTP}_i = \alpha + \beta \text{Cue}_i^{\text{pre}} + \gamma' X_i + \varepsilon_i, \quad (1)$$

by OLS on the Week 2 cross-section, where WTP_i is the last-switch WTP measure (in GBP) constructed in Appendix E, X_i is a vector of individual-level covariates grouped by socio-demographic, behavioural and shopping preferences.²¹ ε_i is the error term clustered at the participant level. The coefficient of interest, β , is the average treatment effect on WTP of receiving the identity cue immediately before the WTP elicitation.

The estimation sample retains participants with at most three preference reversals in the MPL ($n_{\text{changes}} \leq 3$, 97% of the analysis sample; see Appendix E and Table A.14). As a robustness check, we re-estimate equation (1) on the strict single-switcher and non-switcher subsample ($n_{\text{changes}} \leq 1$). Results are reported in Section 6 and Appendix A.

5.2 Basket Emissions Estimation Strategy

The panel basket-emissions analysis starts from the 828 participants with at most one MPL preference switch ($n_{\text{changes}} \leq 1$) and the main IPW-DiD specification further trims the consistent corner cases at $|\text{WTP}| = £7$, leaving an estimation sample of

²¹ Covariates include Gender, Education, Children, Income, Age, Green Identity, Search Effort, Dietary Restrictions, Items Purchased and Checkout Spend.

778 participants.

For the effect of the sorting layout on basket emissions, we use a log-linear difference-in-difference specification with individual fixed effects on $\ln(\text{CO}_{2e,it})$ and the two-period panel.²² The treatment coefficient captures $\ln(\text{CO}_{2e,i,\text{Wk2}}/\text{CO}_{2e,i,\text{Wk1}})$, the proportional within-person change in basket emissions.²³ We estimate the following regression:

$$\ln(\text{CO}_{2e,it}) = \alpha_i + \lambda \text{Post}_t + \tau (\text{Eco Sorting}_i \times \text{Post}_t) + \delta' Z_{it} + u_{it}, \quad (2)$$

We report the coefficient of interest τ both as a log coefficient and converted to a percentage change via $(e^\tau - 1) \times 100$ (Puhani, 2012), and rescaled to kg CO₂e using the Week 2 standard-sorting mean. The vector Z_{it} collects time-varying covariates, while the time-invariant variables are absorbed by the individual fixed effects.

To recover the causal effect of the sorting layout on basket emissions, we follow Epperson and Gerster (2026) and implement an inverse-probability-weighted (IPW) DiD specification that uses the WTP as the propensity score (Epperson and Gerster, 2026; Hirano et al., 2003; Imbens and Wooldridge, 2009; Imbens, 2024; Imbens and Xu, 2025). Although the MPL row draw is itself randomised, the realised Eco Sorting_i indicator is not unconditionally independent of potential outcomes, it is a function of the participant's stated WTP w_i . Participants with higher (lower) w_i are mechanically more (less) likely to land in the eco-sorted layout. Unconfoundedness of layout assignment with potential basket emissions therefore holds only conditional on WTP, $\{Y_{it}(0), Y_{it}(1)\} \perp \text{Eco Sorting}_i \mid w_i$ (Epperson and Gerster, 2026; Rubin, 1974).

²²The log specification was included in the pre-registered analysis plan. Distributional statistics for the within-person change in both raw and log form are reported in Figure A.6.

²³With $T = 2$, the FE estimator is numerically equivalent to first-differencing.

We re-estimate equation (2) weighting by inverse-propensity weights, with the WTP-conditional layout assignment probability as the design-known propensity score:

$$\hat{p}(w_i) \equiv \Pr(\text{Eco Sorting}_i = 1 \mid w_i) = \sum_{r: \text{eco}(w_i, r) = 1} \pi_r, \quad (3)$$

where π_r are the pre-specified, unequal MPL row-draw probabilities described in Section 2.3 (18/53 at each of the two corner rows, 1/53 at each of the seventeen interior rows). By construction, $\hat{p}(w_i)$ is a step function that is monotonically increasing in w_i , with two equal-sized jumps of $18/53 \approx 0.34$ at the corners (from $w_i = -7$ to $w_i = -5$ and from $w_i = 5$ to $w_i = 7$) and uniform increments of $1/53$ across the interior support; the symmetric corner mass therefore enters the propensity score symmetrically in w_i , despite the two-sided range of WTP values. Table A.16 reports worked examples of $\hat{p}(w_i)$ and the implied ATT weights for representative WTP values, including the trimmed corner cases.

Given the design-known propensity score $\hat{p}(w_i)$, we estimate the average treatment effect of eco-sorting on basket emissions using the inverse-probability-weighting (IPW) estimator of Imbens and Wooldridge (2009) (see also Imbens (2024); Epperson and Gerster (2026)):

$$\hat{\tau}_{\text{IPW}} = \frac{1}{n} \sum_{i=1}^n \left[\frac{\text{Eco Sorting}_i \Delta Y_i}{\hat{p}(w_i)} - \frac{(1 - \text{Eco Sorting}_i) \Delta Y_i}{1 - \hat{p}(w_i)} \right], \quad (4)$$

where $\Delta Y_i = \ln(\text{CO}_{2e,i,\text{Wk2}}) - \ln(\text{CO}_{2e,i,\text{Wk1}})$ is the within-person change in log basket emissions, so that equation (4) delivers the IPW analogue of the DiD coefficient τ in equation (2). Each treated observation is reweighted by $1/\hat{p}(w_i)$ and each control observation by $1/[1 - \hat{p}(w_i)]$, the inverse probability of (not) receiving the realised layout assignment given w_i . In practice we implement equation (4) as a weighted version of the DiD regression (2), using ATT weights equal to 1 for participants assigned to the eco-sorted layout and $\hat{p}(w_i)/[1 - \hat{p}(w_i)]$ for those assigned to

the standard layout, normalised within arm (see Table A.16 for worked examples of the resulting weights at representative WTP values).

We trim only the exact corner values $w_i = \pm£7$, where $\hat{p}(w_i) \in \{0, 1\}$ and the layout assignment is non-random by design: these participants choose the same layout in every MPL row, so the random row draw does not generate variation in their assigned layout (see Table A.16, corner rows). This trimming is symmetric in w_i , follows directly from the support of $\hat{p}(w_i)$, and retains all interior WTP values where identification is feasible, following the same minimal trimming rule as Epperson and Gerster (2026). Overlap of the propensity score on the remaining support is verified via kernel-density plots of $\hat{p}(w_i)$ by realised layout arm (Appendix A). As a robustness check, we re-estimate the IPW-DiD using empirical bin probabilities $\tilde{p}(w_i)$ (the share of participants in each WTP bin actually assigned to the eco layout), following Hirano et al. (2003) who show that estimated propensities can dominate known propensities in finite samples even when the latter are correct by design.

To test whether the identity cue exposure on the eco-sorting effect (H2c), we additionally estimate a version of equation (2) interacting $\text{Eco Sorting}_i \times \text{Post}_t$ with those that received the cue ($\text{Cue}_i^{\text{pre}}$ and $\text{Cue}_i^{\text{post}}$ arms). The cue main effects on the basket emissions serve as a mechanism check: the pre-registered prediction is that the direct cue effect on basket emissions is small or null, with the cue acting on baskets only indirectly through its effect on WTP and hence on layout assignment. Finally, motivated by the WTP distribution, we report CATEs by WTP type, namely *avoiders* ($w_i < 0$), *indifferent* ($w_i = 0$), and *seekers* ($0 < w_i < 7$), to probe possible willful-ignorance and preference-heterogeneity patterns (Saccardo and Serra-Garcia, 2023; Epperson and Gerster, 2026).

6 Results

In this section, we first present descriptive evidence on baseline shopping behaviour and demand for the eco-sorted layout (Section 6.1). We then report the pre-registered results on the effect of the identity cue on WTP (Section 6.2) and the effect of the eco-sorted layout on basket emissions (Section 6.3). Heterogeneity analyses and robustness checks are reported in the relevant subsections and in Appendix A.

6.1 Descriptive Statistics and Demand for Eco Sorting

6.1.1 Baseline Shopping Behaviour

Our sample composition is described in Section 3, and balance across treatment arms is verified in Table 2. Here we focus on Week 1 shopping behaviour, before any treatment is introduced. Full Week 1 descriptive statistics are reported in Table A.1.

Week 1 behaviour is consistent with participants engaging seriously with the shopping task. Mean (and median) basket spend is approximately £40 with a standard deviation just under £9, well above the enforced £25 lower bound and close to the £50 ceiling, despite the binding budget constraint and the incentive-compatible lottery in which winners receive their realised basket plus any unspent budget.²⁴ Moreover, participants purchase a median (mean) of 18 (19) items with a standard deviation of 6, well above the minimum required to clear the budget floor while including the six prescribed categories described in Section 2.3.²⁵

Search effort, measured as time spent on the shopping task, averages 11.2 minutes (median 9.8; s.d. 14.4), and is right-skewed. Week 1 basket emissions average 7.3 kg CO₂e (median 6.6; s.d. 4.0; Figure A.5) and exhibit the right skew that is characteristic of grocery emissions in prior food research (Panzone et al., 2021b;

²⁴Participants could not check out if basket spend fell outside the £25–£50 range.

²⁵Participants were not informed that completion of the prescribed list was non-binding for check-out.

Clark et al., 2022; Poore and Nemecek, 2018).

6.1.2 Baseline demand for Eco-Sorting Layout

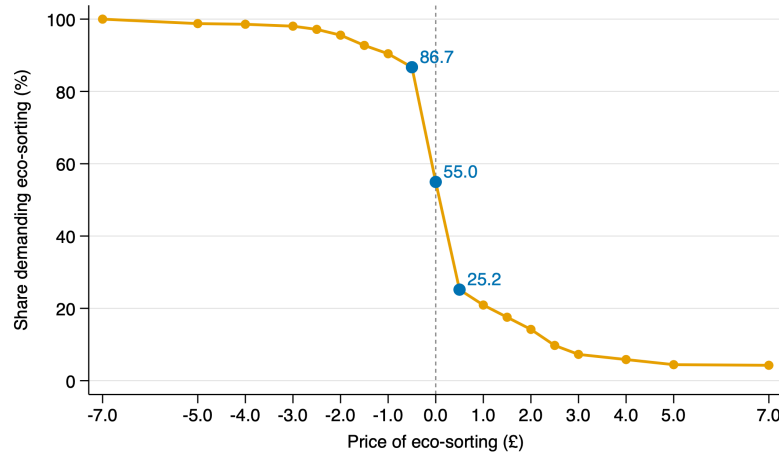


Figure 2: Demand for Eco-sorting Layout

Note: The figure plots the percentage of the sample willing to have the eco-sorting at each given price. Sample restricted to the control and post-MPL cue arms ($N = 546$), and those with 3 or fewer preference reversals in the MPL. Figure markers show the demand share at $-\text{£}0.50$, $\text{£}0$, and $\text{£}0.50$.

Figure 2 presents the baseline demand curve for the eco-sorted layout among participants who were not exposed to the identity cue before the WTP elicitation (the No-Cue and Post-MPL Cue arms). Demand decreases from 100% at $-\text{£}7$ (i.e. the participant is paid $\text{£}7$ to accept the eco-sorting layout) to 4.4% at $\text{£}7$.

The figure reveals two features of demand for the eco-sorted layout. First, eco-sorting avoidance is non-trivial. At $\text{£}0$, only 54.8% of the sample are willing to have the eco-sorting layout.²⁶ Second, demand drops sharply around zero, with a much larger change than at higher or lower prices.²⁷ When the eco-sorting price

²⁶At $\text{£}0$ the sorting layout, unlike information, is not entirely costless to the consumer, who might anticipate extra time searching for relatively less green items.

²⁷The concentration of WTP around $\pm\text{£}1$ also validates our quasi-random design, which draws with much higher probability from MPL rows associated with $|\text{WTP}| > \text{£}5$.

shifts from -£0.50 (where the participant is paid to accept it) to £0, demand falls by 32 percentage points (from 86.8% to 54.8%); a further 30 percentage point fall occurs when the price moves from free to £0.50 (from 54.8% to 24.4%). Demand is therefore highly elastic around zero, and there is substantial potential to shift demand from small movements in consumer WTP. This pattern echoes the high price elasticity around zero documented for information avoidance (Epperson and Gerster, 2026; Saccardo and Serra-Garcia, 2023; Momsen and Ohndorf, 2020; Serra-Garcia and Szech, 2022). We test below whether an information cue can shift WTP for the eco-sorting layout.

Table A.3 in Appendix A reports the socio-demographic, preference and shopping-behaviour characteristics of seekers and avoiders of the eco-sorting layout, defined as those with WTP above or below zero. Seekers and avoiders differ on essentially two dimensions: seekers are more likely university educated (71% vs. 61%, significant at the 5% level) and report a higher pre-intervention (Week 1) green identity, consistent with willingness-to-pay for a greener shopping experience. Across all other characteristics, age, gender, familial situation (with children), income group, Week 1 grocery spend, number of items bought, time spent shopping, whether they shop for others, and even Week 1 basket emissions (despite eco labels already being visible), we detect no significant differences at the 5% level. The same pattern holds when restricting to the marginal population on the demand curve, where only green identity matters (WTP between -£1 and £1; see Table A.4).

This marginal mass of 62% of participants (those with WTP in $[-£1, £1]$, whose preference for eco-sorting switches around £0) and who differ from the rest of the sample only in green identity is precisely the population the identity cue is designed to move. A small, costless cue that nudges these marginal consumers across the zero-price threshold can therefore translate into a disproportionately large shift in demand for eco-sorting, and, through layout assignment, into a meaningful change

in basket-level environmental impact.

6.2 Effect of Identity Cue on Eco-Sorting WTP

6.2.1 Average treatment effects on the treated

Table 3 reports estimates of the effect of the identity cue on WTP for the eco-sorted layout (equation (1)). Building from a simple bivariate regression to specifications including socio-demographic factors (gender, age, income, education), shopping behaviour (week 1 budget, number of items bought, time spent shopping) and green preferences (self-reported green identity in week 1) as controls, we observe a consistent and statistically significant positive effect of the pre-MPL identity cue on WTP for the eco-sorting layout. The point estimate of the cue effect is around £0.45. Given the high elasticity of demand around zero shown in Figure 2, this is a substantial shift relative to the baseline mean WTP. Figure A.1 corroborates this with the unconditional distribution of WTP, showing a clear rightward shift for the pre-MPL cue arm relative to the rest of the sample.

Table A.5 reports the full set of control variable estimates. As noted above, no socio-demographic or shopping-behaviour variable significantly affects WTP for the eco-sorting layout. Participants with low self-reported green identity in week 1 have a lower WTP for the eco-sorting layout (roughly -£0.61), and those with a higher number of items bought in week 1 have a higher WTP (+£0.48).

6.2.2 Identity Updating from the Identity Cue

The assumed mechanism is that the identity cue updates the consumer's self-image as a green consumer (Bénabou and Tirole, 2011; Tonke, 2025), which in turn drives the shift in WTP for the eco-sorting layout. In Appendix Section A.3.1 we consider this mechanism. We observe a positive and significant increase in self-reported green identity from week 1 to week 2 for participants who received the

Table 3: Effect of the pre-MPL identity cue on WTP for eco-sorting.

	Simple	+ Socio	Full controls	Full controls w/ Multiple-switchers
Pre-MPL cue	0.404*** (0.156)	0.416*** (0.155)	0.453*** (0.154)	0.457*** (0.152)
Socio-demographics	No	Yes	Yes	Yes
Green ID (Wk1)	No	No	Yes	Yes
Shopping prefs.	No	No	Yes	Yes
Participants	827	827	827	849
Adj. R ²	0.007	0.009	0.048	0.046

Notes: Columns 1-3 include only those with 1 or 0 preference reversals in the MPL. Column 4 uses the sample of those with 3 or fewer preference reversals. OLS with cluster-robust standard errors (id-level) in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

cue (whether pre- or post-MPL), of 0.831 points on our 3–33 green identity scale (Table A.3.1).²⁸ While the point estimates suggest a somewhat larger increase for the pre-MPL cue arm (around 1.1–1.2 points, compared to 0.5–0.7 for the post-MPL cue), the difference between the two cue arms is not statistically significant (Wald test, SE=0.38, $p=0.25$, Appendix A.3.1), and the significance of the post-MPL cue varies with the sample restriction (Table A.3.1). Overall, we do observe an effect of the cue on self-reported green identity, but lack the power to fully resolve the question of cue timing conclusively.

6.2.3 Shape of the Identity Cue Effect

Following the pre-registered pre-analysis plan, we also examine whether the identity cue has heterogeneous effects based on baseline green identity, notably the inverted-U shape suggested by the Bénabou and Tirole (2011) identity model. Under that model, only those with moderate green identity in week 1 should be pos-

²⁸Based on a DiD specification with individual fixed effects and clustered standard errors. The green identity measure was collected post-grocery shop each week and any updating is independent of sorting layout in the analysis.

itively influenced by a cue suggesting they are green consumers,²⁹ Table A.6 and Figure A.2 report the heterogeneity analysis. Table A.6 tests the interaction between the pre-MPL cue and the tercile version of the baseline green identity variable. Although the point-estimate signs follow the inverted-U pattern, the interaction is not statistically significant.³⁰ Testing the interaction with a continuous and quadratic version of the baseline green identity variable also yields the correct sign but no statistical significance. From Figure A.2 we observe that pointwise marginal effects of the cue along the continuous green identity variable are positive and significant only at moderate green identity levels, consistent with the inverted-U pattern, although the confidence bands widen at the tails due to smaller sample size. Overall, while the qualitative shape is consistent with the model, we find no conclusive evidence of heterogeneity in the cue effect by baseline green identity at the typical significance levels.

6.3 Effect of Eco-sorting on Basket Emissions

6.3.1 Average Treatment Effects on Basket Emissions

The eco-sorted layout is predicted to reduce basket emissions by shifting which products are encountered first. Table 4 reports the ATE of the eco-sorted layout on basket emissions, estimated using the log-linear difference-in-differences specification with individual and week fixed effects (TWFE) described in Section 5.2. Our preferred estimate is from the IPW analysis sample, restricted to single-switchers ($n_{\text{changes}} \leq 1$) and excluding corner non-switchers ($\hat{p}(w_i) \in (0, 1)$). For completeness, we also report estimates on the full sample and on the IPW sample without IPW weights (columns 1 and 2 of Table 4). The eco-sorting layout produces a statistically significant reduction in basket emissions, with a point estimate of ap-

²⁹The model implies that the low and high green identity groups would ignore the message, not updating their identity beliefs nor their subsequent behaviour, believing either that they are already green or that they are not green at all.

³⁰This test was run on the sample with up to 1 and up to 3 preference reversals.

proximately -0.30 , which under the Puhani (2012) conversion corresponds to a $(e^{-0.30} - 1) \times 100 \approx 26\%$ reduction (visualised in panel 1 of Figure 3). The estimate is robust and significant across specifications, indicating that neither the IPW weighting nor the sample restriction materially affects the estimated effect.³¹ In absolute terms, this corresponds to a reduction of approximately 1.93 kg CO₂e per basket, relative to a week 1 (pre-intervention) sample mean of 7.3 kg (Appendix Section C.1), a meaningful reduction in the emissions footprint of weekly grocery shopping.

Table 4: Effect of eco-sorting on basket CO₂e: sample restriction vs IPW reweighting

	Full sample (unweighted)	IPW sample (unweighted)	IPW sample (IPW-weighted)
Eco-sorting $\times$ Post	-0.307*** (0.040)	-0.309*** (0.042)	-0.304*** (0.043)
Observations	1750	1556	1556
% change (Puhani)	-26.4	-26.6	-26.2
Participants	875	778	778

Notes: Two-way fixed-effects (week and individual) DiD on log basket CO₂e. Cluster-robust SEs at the participant level (in parenthesis). Model 1: full panel, no MPL coherence screen, no propensity trim. Model 2: single-switchers, excl. corner non-switchers: $\hat{p}(w_i) \in (0, 1)$ Model 3: same sample as (2), reweighted with IPW weights Percent change via Puhani (2012): $(\exp(\delta) - 1)$ times 100. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

6.3.2 Mechanism: Search Effort

The primary mechanism through which product ordering affects consumer choice is search costs. Placing greener items in prime positions lowers the effort required to find them, making it more likely that a satisficing consumer adds them to the basket before the search-effort budget is exhausted (Simon, 1955; Derakhshan et al., 2022; Johnson et al., 2012). We test this mechanism directly using a pre-registered

³¹The identity-cue indicator is time-invariant and absorbed by the individual fixed effects; with the IPW weighting, identity cue and layout assignments are both independently randomised.

heterogeneity analysis that interacts the eco-sorted layout with Week 1 search effort, defined as a binary split at the median shopping time on the platform.

Panel 2 of Figure 3 present CATEs by Week 1 search effort. The lower search effort group shows a much larger reduction in basket emissions, with an estimated reduction of around 38% compared to 12% for the higher search effort group. The difference is statistically significant at the 1% level ($p\text{-value} < 0.001$), consistent with the search-cost mechanism. Consumers who invest less search effort are most exposed to whatever the layout places first, and therefore most responsive to the eco-sorted ordering.

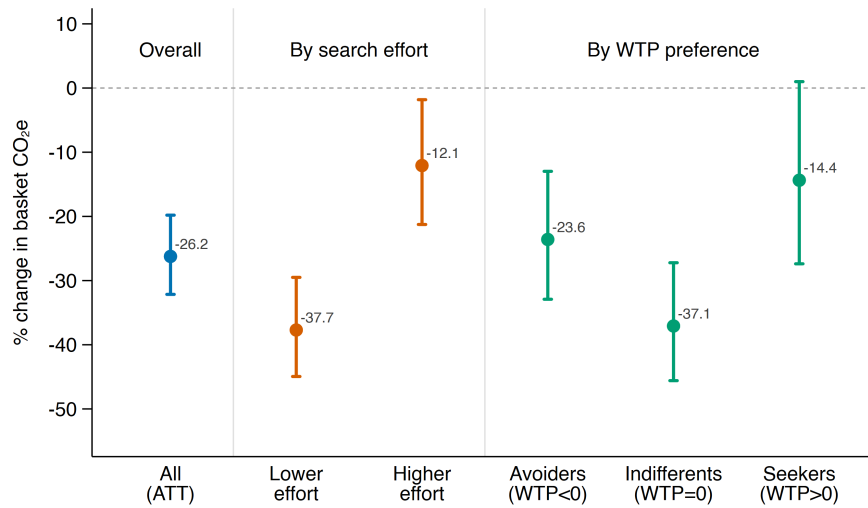


Figure 3: Eco-sorting effect on basket emissions: overall, by WTP preference type, and by search effort

Note: IPW FE-DiD estimates of the eco-sorting effect on basket CO₂e, computed via Puhani's conversion as $(\exp(\delta) - 1) \times 100$, where δ are the log coefficients on eco-sorting. 95% CIs. Sample: IPW analysis sample (single switchers, excl. corner non-switchers ($\hat{p}(w_i) \in (0, 1)$), 778). Panel 1 presents the overall ATT from Table 4. Panel 2 (middle) presents CATEs by above vs below median Week 1 search effort; $p < 0.001$ for difference by search effort. Panel 3 plots CATEs by ex-ante WTP preference type from Table 5; Avoiders: WTP < 0; Indifferent: WTP = 0; Seekers: WTP > 0.

6.3.3 Effect Heterogeneity by Layout Preference

A related source of heterogeneity is participants' ex-ante preference for the eco-sorted layout. Consumers who actively sought the layout (Seekers) are likely already more environmentally conscious shoppers, potentially leaving less room for the layout to shift their behaviour relative to those who are indifferent or resistant to it. We estimate the conditional average treatment effect (CATE) by partitioning the IPW sample into three groups defined by participants' eco-sorting preference prior to assignment: Avoiders ($WTP < 0$), Indifferent ($WTP = 0$), and Seekers ($WTP > 0$), and estimate the IPW DiD specification separately for each group.

The third panel of Figure 3 plots the CATEs by ex-ante WTP preference type. The eco-sorted layout reduces emissions most for the Indifferent group, with an estimated reduction of 37%, followed by Avoiders at roughly 24%; both effects are significant at the 1% level. For the Indifferent group this corresponds to a saving of more than 2 kg CO₂e per shop. Full regression results are reported in Table 5. Seekers are the least affected, plausibly because higher WTP is positively associated with environmental behaviour, leaving less room for reduction relative to their already lower week 1 baskets (their reduction is significantly smaller than the Indifferent group, Wald test, $p = 0.006$). The Seekers point estimate is around 14% and is significant only at the 10% level, though we note the smaller sample of 182, against 339 Avoiders and 257 Indifferent.

We therefore find heterogeneity in the eco-sorting effect across ex-ante preference types (Wald test, $p = 0.018$, Table 5). Crucially, the Avoiders' effect is not statistically distinguishable from the Indifferent or Seeker effects (Wald tests, $p = 0.051$ and $p = 0.287$, respectively), providing no evidence of reactance or resistance among those who would have preferred to avoid the layout. The pattern of heterogeneity is consistent with less scope for emissions reduction among consumers with higher ex-ante WTP.

Robustness checks are reported in Appendix Section A.4.1 (Tables A.10 and A.11), re-estimating the CATEs on the full sample, without IPW weighting, and using the doubly-robust adjusted IPW estimator of Sant’Anna and Zhao (2020). Estimates and inference are robust, with the note that the Seekers point estimate becomes significant at the 1% level in the less restricted sample while remaining similar in magnitude across specifications.

Table 5: CATE of eco-sorting on basket CO₂e by WTP preference type

	Avoiders	Indifferent	Seekers
Eco-sorting $\times$ Post	-0.269*** (0.066)	-0.463*** (0.074)	-0.155* (0.084)
Observations	678	514	364
% change (Puhani)	-23.6	-37.1	-14.4
Participants	339	257	182

Notes: Two-way fixed-effects (week and individual) DiD on log basket CO₂e. Sample: single-switchers, excl. corner non-switchers ($\hat{p}(w_i) \in (0,1)$). The estimates are with the IPW weighting. Avoiders: WTP < 0; Indifferent: WTP = 0; Seekers: 0 < WTP < 7. Cluster-robust SEs at the participant level (in parenthesis). Joint test that sorting effect varies across the WTP types: $p = 0.018$. Percent change via Puhani (2012): $(\exp(\delta) - 1) \times 100$. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

6.4 Complementarity at the Point of Choice

The preceding sections establish that the eco-sorted layout reduces basket emissions through search-cost and primacy effects, with the largest effects among Indifferent and low-search-effort participants. A further pre-registered question is whether the identity cue amplifies this effect at the point of choice, or whether the two instruments operate independently. We test this across three outcomes: basket emissions (H2d), green identity updating, and consumer satisfaction.

6.4.1 Basket Emissions

We estimate the IPW FE-DiD specification of equation (2) interacted with the identity-cue indicator (combining the Pre- and Post-MPL Cue arms).³² The left panel of Figure 4 reports the estimated emissions effect of eco-sorting separately for cued and non-cued participants. The two effects are essentially identical with a 26.5% reduction with the cue and a 26.0% reduction without it, with no statistical difference in the three-way interaction ($p = 0.949$). The identity cue does not amplify the eco-sorting effect on basket emissions.

One interpretation of this null is that the layout has already captured most of the available within-shop emissions response. The largest CATEs are for Indifferent and low-search-effort participants (Table 5, Figure 3), suggesting that the dominant mechanism is the search-cost effect of the layout itself, leaving little additional behavioural margin for the cue.

Together with the WTP results in Section 6.2, these findings suggest that the two instruments operate on distinct aspects of the decision process. The cue shifts demand for the layout ex-ante (H1a), while the layout drives the in-shop behavioural change (H2a), with no detectable complementarity at the point of choice.

6.4.2 Green Identity

A second margin on which the two interventions could interact is self-image. If being assigned to the eco-sorted layout (and shopping greener as a result) reinforces the cue's identity message, we would expect a larger week-1-to-week-2 increase in self-reported green identity for participants who received both. We test this in Appendix Section A.3.2 by extending the panel fixed-effects DiD specification used to estimate the cue's effect on green identity (Appendix Section A.3.1) to additionally

³²We do not have power to also split this interaction by WTP sign, which would require dividing the sample further between the Pre-MPL cue (where the cue affects WTP) and the Post-MPL cue arms.

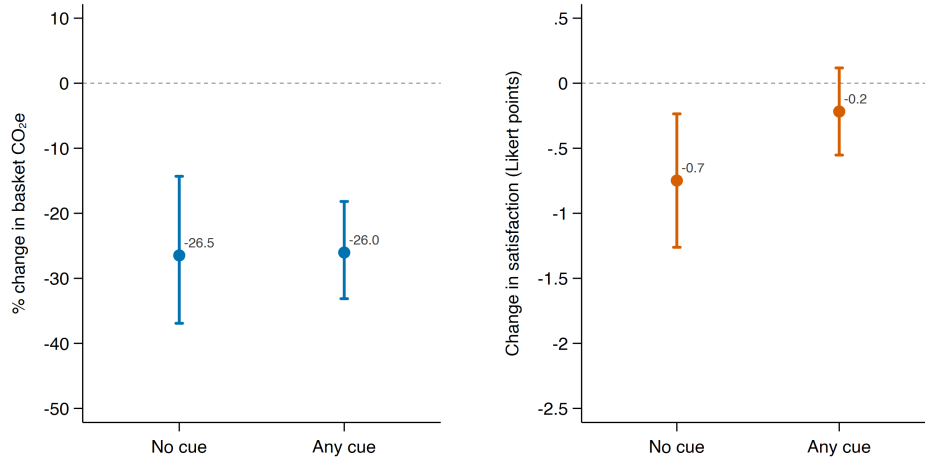


Figure 4: Impact of eco-sorting by cue exposure on basket emissions and satisfaction

Note: IPW FE-DiD estimates of the eco-sorting effect by cue exposure, recovered from a saturated specification interacting eco-sorting with the cue indicator. 95% CIs. Sample: IPW analysis sample (single switchers, excl. corner non-switchers ($\hat{p}(w_i) \in (0, 1)$), 778). *Left panel:* ATE in percentage change in basket CO₂e, computed via Puhani's conversion as $(\exp(\delta) - 1) \times 100$; $p = 0.949$ for the cue interaction. *Right panel:* ATE in change in consumer satisfaction (Likert points, scale 0–10); $p = 0.089$ for the cue interaction.

include the eco-sorting indicator and its interaction with the post-treatment period. Across all specifications, the eco-sorting effect on green identity is small and not statistically significant, and the three-way cue-by-layout interaction is close to zero, though we note that the triple interaction specification is underpowered. The eco-sorted layout reduces basket emissions without updating self-reported green identity, consistent with it acting through the structure of the choice environment rather than through the kind of salient, diagnostic act that feeds back into self-image (Bénabou and Tirole, 2011; Bordalo et al., 2020). The eco-sorted layout is, in this sense, identity-neutral.

6.4.3 Consumer Satisfaction

A final margin we examine is whether the identity cue affects the shopping experience of participants assigned to the eco-sorted layout. We re-estimate the cue-by-layout interaction of equation (2), with self-reported shopping satisfaction (a 0 to 10 Likert scale) as the dependent variable, measured in both weeks directly after participants completed their shop. The right panel of Figure 4 shows that assignment to the eco-sorted layout reduces satisfaction by 0.7 Likert points in the absence of an identity cue. Among participants who received the cue, the reduction is small (-0.2) and not statistically significant, with the difference between the two groups significant at the 10% level ($p = 0.089$). The identity cue therefore appears to offset the satisfaction cost of the eco-sorted layout, consistent with participants finding it more acceptable when their environmental self-image has been activated.³³

7 Managerial and Policy Implications

The preceding sections establish the causal effects of each instrument and their interaction. This section considers what they imply for policymakers and retailers. Specifically, whether the combination produces unintended costs, who benefits and who does not, and whether the intervention design is ethically defensible. Supporting results are in Appendix B.

7.1 Unintended Effects on Spend, Time, and Satisfaction

Two potential concerns accompany any product-ordering intervention. For consumers, eco-sorted layouts may increase total basket spend if greener products carry

³³The cued and non-cued arms are well balanced on WTP composition in the IPW sample by design, with no significant differences in avoider, indifferent or seeker (mean WTP test $p = 0.91$ unweighted, $p \geq 0.99$ IPW-weighted), so the cue's effect on satisfaction is not driven by differential preference composition across arms.

a price premium, raising the cost of the weekly shop, and may increase the time spent searching for preferred items if greener alternatives are poor substitutes for those displaced lower in the list. For retailers, higher spend would benefit revenue but longer shopping times and lower satisfaction could reduce platform engagement and repeat usage. We examine each in turn.

Table A.12 in Appendix B reports IPW FE-DiD estimates of the effect of the eco-sorted layout and the identity cue on total basket spend and shopping time, using the same specification as for the emissions outcome and controlling for the number of items purchased. Column 1 shows that the eco-sorted layout is associated with a decrease in basket spend of around £2.50 relative to the standard layout, providing no evidence of a green price premium in this setting. Column 3 shows no statistically significant effect on shopping time. For both outcomes, the identity cue produces no additional interaction effect (Columns 2 and 4). Both spend and shopping time are, as expected, positively related to the number of items purchased.

On satisfaction, the eco-sorted layout produces a small reduction among non-cued participants (less than 1 Likert point), but this reduction is not statistically significant when the layout is paired with the identity cue (see Section 6.4). The identity cue therefore improves the palatability of the layout without amplifying its emissions effect.

7.2 Distribution of Effects

Beyond the heterogeneity by WTP type and search effort reported in Section 6.3, we examine whether the eco-sorted layout affects some socio-demographic groups disproportionately. Figure A.3 reports IPW FE-DiD estimates of the layout effect by age, gender, income, and education.

We find no evidence of heterogeneity by gender or education (joint tests $p = 0.618$

and $p = 0.920$). The eco-sorting effect is slightly decreasing in income (across three income groups, less than £30k, £30k-£50k, more than £50k per annum) but this is not statistically significant (joint test $p = 0.276$). Finally, we find no statistically significant heterogeneity by age (joint test $p = 0.308$), although the 32-43 age group shows largest point-estimate reduction in emissions from the eco-sorting layout.

7.3 Ethical Intervention Design

A guiding principle of behavioural policy design is that interventions should preserve individual agency and freedom of choice, remain transparent and avoid deception (Thaler, 2018; Thaler and Sunstein, 2009; Lades and Delaney, 2022). We briefly evaluate the cue-plus-eco-sorting combination against the seven dimensions of the FORGOOD framework (Lades and Delaney, 2022): Fairness, Openness, Respect for autonomy, Goals, Opinions, Options, and Delegation.

On Fairness, the combined intervention does not disproportionately affect consumers along socio-demographic lines, imposes no detectable cost in time or expenditure, and both search-effort groups benefit from the layout. On Openness and Respect for autonomy, the identity cue is an informational statement rather than a directive and could easily be ignored; the eco-sorted layout changes the order in which products are presented but does not alter prices, products, or the underlying choice set. On Goals and Opinions, the intervention targets a pre-registered environmental objective and combines a preference-activating instrument with a structural one. On Options and Delegation, the eco-sorted layout is offered alongside existing sorting options.

7.4 External Validity

Finally, we assess external validity against the SANS framework (Al-Ubaydli et al., 2017; List, 2020). On Selection, a sample of UK participants was recruited via Prolific with eligibility filters selected to match the target population (primary household shoppers with prior online grocery experience). Layout assignment is quasi-random by design, with WTP-conditional non-randomness corrected by inverse probability weighting. Attrition between weeks is independent of treatment by design, and balance on observables is verified in Table A.2. On Naturalness, the study is a framed field experiment (Harrison and List, 2004) on a platform replicating UK online grocery retailers, creating a familiar shopping environment with a complete product catalogue (9,500 real products). Moreover, participants faced real monetary incentives for both the WTP elicitation and the shopping task making outcomes realised purchases rather than stated intentions. On Scaling, the identity cue is informational rather than prescriptive, linking a widely desirable self-image to a broadly available behaviour, and does not mention the eco-sorted layout or prescribe specific items. The observed shift in WTP for an unmentioned layout strengthens the case for the cue’s effectiveness as a spillover from identity to related behaviour, consistent with Tonke (2025). As with most studies of this kind, the analysis is partial equilibrium. We observe demand-side behavioural responses but not how retailers or producers would respond at scale.

More broadly, these results illustrate a general principle applicable beyond the grocery context. An identity cue lowers the psychological cost of voluntarily adopting a more effective, identity-consistent choice environment (Bénabou and Tirole, 2011), while the restructured environment drives the behavioural change (Johnson et al., 2012; Thaler and Sunstein, 2009). The cue’s aggregate impact is likely largest in settings where demand for the identity-consistent behaviour is highly elastic around the indifference margin, as we find here: a small shift in motivated self-signalling

moves a disproportionately large share of consumers across the opt-in threshold (Epperson and Gerster, 2026; Serra-Garcia and Szech, 2022). The choice architecture intervention's impact is in turn likely greatest in settings where the structure of the environment is a key driver of decisions, such as online retail or energy consumption, as opposed to high-stakes deliberative decisions where cognitive and motivational factors dominate (e.g. high-stakes financial decisions). The complementarity principle applies wherever a choice architecture intervention is effective but faces adoption friction.

8 Discussion and Conclusion

Belief-based identity models predict that making the link between behaviour and self-image salient can motivate identity-consistent action (Bénabou and Tirole, 2011; Tonke, 2025). In this paper, we test whether this mechanism can resolve a practical tension in intervention design: choice architecture interventions are among the most effective behavioural tools available, but imposing them is contested and voluntary uptake is low. In a pre-registered framed field experiment with 1,087 UK participants, we find that an identity cue raises willingness to pay for an eco-sorted product layout by approximately £0.45, a modest shift that moves a disproportionately large share of consumers across the opt-in margin given the high elasticity of demand around zero. The eco-sorted layout, in turn, reduces basket emissions by around 26%, consistent with a search-cost mechanism. The two effects operate through separate channels whereby the cue shifts the decision to adopt the architecture, but has no direct effect on the choices made within it.

Our results speak to the conditions under which the identity-behaviour feedback loop closes (Charness and Chen, 2020). Extending prior evidence, in which self-perception is measured cross-sectionally after the intervention (Tonke, 2025), we observe the full chain within the same individuals where green identity is elicited both before and after the cue, and the within-person difference-in-differences shows the cue raising self-reported green identity by 0.83 points on the 3–33 scale. This is a belief update, reinforcing the post-intervention level difference (Tonke, 2025) consistent with the belief-updating mechanism of Bénabou and Tirole (2011), and this belief update translates into higher willingness to pay for the eco-sorted layout. The opt-in layout choice is a single, salient act experienced as a diagnostic commitment about the self, and motivated self-signalling (Bénabou and Tirole, 2016; Bénabou, 2015) can reach it. The loop does not, however, extend to basket-level behaviour directly. Individual item selections within the shop are

shaped primarily by what products are encountered, in what order, and at what cognitive cost, and the identity cue's signal is crowded out at that stage. The salience and diagnostic value of the decision is therefore a key moderator of whether identity translates into behaviour (Bordalo et al., 2020).

The cue's contribution to aggregate emissions savings operates through uptake composition. By shifting participants along the WTP distribution toward and beyond the opt-in threshold, it enlarges the population over whom the layout's large effect is realised, without amplifying that effect within individuals. Appendix Figure A.4 provides back-of-the-envelope estimates of the aggregate emissions gain from combining the two instruments relative to deploying the choice architecture intervention alone. The identity cue also offsets the small reduction in shopping satisfaction produced by the eco-sorted layout among those randomly assigned to it, providing further support for the combined intervention.

Our micro-level panel design also allows us to bring direct evidence on a number of further predictions from the belief-based identity literature that, to our knowledge, have so far not been tested within individuals. First, we observe the full belief-to-preference-to-behaviour chain in the same individuals, whereas prior field evidence infers the belief-updating step from post-intervention measurement alone (Tonke, 2025). Second, we test Bénabou and Tirole (2011)'s inverted-U prediction that only moderate-identity types should be sensitive to and update from the identity cue. The point-estimate signs follow the predicted pattern but the interaction is not statistically significant with our sample, which is not sufficiently powered for this question. Third, we find no evidence of reactance among Avoiders randomly assigned to eco-sorting, whose basket emissions still fall by around 24%. Fourth, the asymmetry between the cue updating identity and the layout moving behaviour without feeding back into self-image provides micro-level evidence that the action-to-identity channel is salience-moderated rather than universal (Bénabou and Ti-

role, 2006, 2011; Bordalo et al., 2020). Finally, we document a large demand elasticity for the identity-consistent choice architecture around the indifference margin, a roughly 30-percentage-point demand drop around £0, suggesting high sensitivity to small friction costs but also large behavioural change opportunities from small interventions. This parallels the sharp price-sensitivity of information avoidance (Epperson and Gerster, 2026; Saccardo and Serra-Garcia, 2023; Serra-Garcia and Szech, 2022; Momsen and Ohndorf, 2020), but operating on the choice architecture of goods layout rather than on information provision.

These results also shed light on why information interventions at the point of choice, such as eco-labels and nutritional scores, have produced modest and inconsistent effects in prior food research (Panzone et al., 2021b; Pechey et al., 2022b; Bollinger et al., 2011; Kiesel et al., 2011; Crosetto et al., 2020; Hainmueller et al., 2015). When within-shop item selection is driven by the structure of the choice environment, interventions that require consumers to consciously process additional information at that stage face the same structural forces that crowd out the identity cue. The implication is not that information is ineffective, but that its effectiveness depends on where in the decision process it is applied. More broadly, our results point to a general design principle: pairing an identity-activating intervention with a choice architecture intervention, and using the former to expand uptake of the latter, can achieve broader coverage and more sustained impact while preserving consumer autonomy (Lades and Delaney, 2022; Sunstein, 2016). This holds wherever the structural intervention is effective but the path to adoption creates friction, whether through inertia, preference heterogeneity, or legitimacy concerns about top-down imposition.

References

- Akerlof, G. A. and Kranton, R. E. (2000). Economics and Identity. *The Quarterly Journal of Economics*, 115(3):715–753.
- Al-Ubaydli, O., List, J. A., and Suskind, D. (2020). 2017 klein lecture: The science of using science: Toward an understanding of the threats to scalability. *International Economic Review*, 61(4):1387–1409.
- Al-Ubaydli, O., List, J. A., and Suskind, D. L. (2017). What can we learn from experiments? understanding the threats to the scalability of experimental results. *American Economic Review*, 107(5):282–286.
- Allcott, H. and Rogers, T. (2014). The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *American Economic Review*, 104(10):3003–3037.
- Andersen, S., Harrison, G. W., Lau, M. I., and Rutström, E. E. (2006). Elicitation using multiple price list formats. *Experimental Economics*, 9(4):383–405.
- Anderson, L. R., Freeborn, B. A., McAlvanah, P., and Turscak, A. (2023). Pay every subject or pay only some? *Journal of Risk and Uncertainty*, 66(2):161–188.
- Atkin, D., Colson-Sihra, E., and Shayo, M. (2021). How do we choose our identity? a revealed preference approach using food consumption. *Journal of Political Economy*, 129(4):1193–1251.
- Augenblick, N. and Nicholson, S. (2016). Ballot position, choice fatigue, and voter behaviour. *The Review of Economic Studies*, 83(2):460–480.
- Bénabou, R. (2015). The economics of motivated beliefs. *Revue d’Économie Politique*, 125(5):665–685.
- Bénabou, R. and Tirole, J. (2004). Willpower and personal rules. *Journal of Political Economy*, 112(4):848–886.
- Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, 96(5):1652–1678.
- Bénabou, R. and Tirole, J. (2011). Identity, morals, and taboos: Beliefs as assets. *The Quarterly Journal of Economics*, 126(2):805–855.
- Bénabou, R. and Tirole, J. (2016). Mindful economics: The production, consumption, and value of beliefs. *Journal of Economic Perspectives*, 30(3):141–164.
- Benjamin, D. J., Choi, J. J., and Strickland, A. J. (2010). Social identity and preferences. *American Economic Review*, 100(4):1913–1928.
- Bertrand, M. and Kamenica, E. (2023). Coming apart? cultural distances in the united states over time. *American Economic Journal: Applied Economics*,

15(4):100–141.

- Bollinger, B., Leslie, P., and Sorensen, A. (2011). Calorie posting in chain restaurants. *American Economic Journal: Economic Policy*, 3(1):91–128.
- Bordalo, P., Gennaioli, N., and Shleifer, A. (2013). Salience and consumer choice. *Journal of Political Economy*, 121(5):803–843.
- Bordalo, P., Gennaioli, N., and Shleifer, A. (2020). Memory, attention, and choice. *The Quarterly Journal of Economics*, 135(3):1399–1442.
- Bott, K. M., Cappelen, A. W., Sørensen, E. Ø., and Tungodden, B. (2020). You’ve got mail: A randomized field experiment on tax evasion. *Management Science*, 66(7):2801–2819.
- Bruner, D. M. (2011). Multiple switching behaviour in multiple price lists. *Applied Economics Letters*, 18(5):417–420.
- Bryan, C. J., Walton, G. M., Rogers, T., and Dweck, C. S. (2011). Motivating voter turnout by invoking the self. *Proceedings of the National Academy of Sciences*, 108(31):12653–12656.
- Charness, G. and Chen, Y. (2020). Social identity, group behavior, and teams. *Annual Review of Economics*, 12(1):691–713.
- Clark, M., Springmann, M., Rayner, M., Scarborough, P., Hill, J., Tilman, D., Macdiarmid, J. I., Fanzo, J., Bandy, L., and Harrington, R. A. (2022). Estimating the environmental impacts of 57,000 food products. *Proceedings of the National Academy of Sciences*, 119(33):e2120584119.
- Cohn, A., Fehr, E., and Maréchal, M. A. (2014). Business culture and dishonesty in the banking industry. *Nature*, 516(7529):86–89.
- Cohn, A. and Maréchal, M. A. (2016). Priming in economics. *Current Opinion in Psychology*, 12:17–21.
- Cook, A. J., Kerr, G. N., and Moore, K. (2002). Attitudes and intentions towards purchasing GM food. *Journal of Economic Psychology*, 23(5):557–572.
- Crosetto, P., Lacroix, A., Muller, L., and Ruffieux, B. (2020). Nutritional and economic impact of five alternative front-of-pack nutritional labels: experimental evidence. *European Review of Agricultural Economics*, 47(2):785–818.
- Demarque, C., Charalambides, L., Hilton, D. J., and Waroquier, L. (2015). Nudging sustainable consumption: The use of descriptive norms to promote a minority behavior in a realistic online shopping environment. *Journal of Environmental Psychology*, 43:166–174.
- Derakhshan, M., Golrezaei, N., Manshadi, V., and Mirrokni, V. (2022). Product ranking on online platforms. *Management Science*, 68(6):4024–4041.

- Epperson, R. and Gerster, A. (2026). Willful ignorance and moral behavior. *Management Science*.
- Hainmueller, J., Hiscox, M. J., and Sequeira, S. (2015). Consumer demand for fair trade: Evidence from a multistore field experiment. *Review of Economics and Statistics*, 97(2):242–256.
- Harrison, G. W. and List, J. A. (2004). Field experiments. *Journal of Economic Literature*, 42(4):1009–1055.
- Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.
- Imbens, G. W. (2024). Causal inference in the social sciences. *Annual Review of Statistics and Its Application*, 11.
- Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86.
- Imbens, G. W. and Xu, Y. (2025). Comparing experimental and nonexperimental methods: What lessons have we learned four decades after LaLonde (1986)? *Journal of Economic Perspectives*, 39(4):173–201.
- Jack, B. K., McDermott, K., and Sautmann, A. (2022). Multiple price lists for willingness to pay elicitation. *Journal of Development Economics*, 159:102977.
- Johnson, E. J., Häubl, G., and Keinan, A. (2007). Aspects of endowment: a query theory of value construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(3):461.
- Johnson, E. J., Shu, S. B., Dellaert, B. G., Fox, C., Goldstein, D. G., Häubl, G., Larrick, R. P., Payne, J. W., Peters, E., Schkade, D., et al. (2012). Beyond nudges: Tools of a choice architecture. *Marketing Letters*, 23(2):487–504.
- Kamenica, E. (2012). Behavioral Economics and Psychology of Incentives. *Annual Review of Economics*, 4(1):427–452.
- Kanay, A., Hilton, D., Charalambides, L., Corrégé, J.-B., Inaudi, E., Waroquier, L., and Cézéra, S. (2021). Making the carbon basket count: Goal setting promotes sustainable consumption in a simulated online supermarket. *Journal of Economic Psychology*, 83:102348.
- Kessler, J. B. and Milkman, K. L. (2018). Identity in charitable giving. *Management Science*, 64(2):845–859.
- Kiesel, K., McCluskey, J. J., and Villas-Boas, S. B. (2011). Nutritional labeling and consumer choices. *Annual Review Resource Economics*, 3(1):141–158.
- Lades, L. K. and Delaney, L. (2022). Nudge FORGOOD. *Behavioural Public Policy*,

6(1):75–94.

- List, J. A. (2011). Why economists should conduct field experiments and 14 tips for pulling one off. *Journal of Economic Perspectives*, 25(3):3–16.
- List, J. A. (2020). Non est disputandum de generalizability? a glimpse into the external validity trial. Technical report, National Bureau of Economic Research.
- List, J. A. and Price, M. K. (2016). The use of field experiments in environmental and resource economics. *Review of Environmental Economics and Policy*, 10(2):206–225.
- Lohmann, P. M., Gsottbauer, E., Farrington, J., Human, S., and Reisch, L. A. (2024). Choice architecture promotes sustainable choices in online food-delivery apps. *PNAS nexus*, 3(10):pgae422.
- Momsen, K. and Ohndorf, M. (2020). When do people exploit moral wiggle room? an experimental analysis of information avoidance in a market setup. *Ecological Economics*, 169:106479.
- Nardotto, M. and Sequeira, S. (2021). Identity, media and consumer behavior. Technical report, CEPR Discussion Paper No. DP15765.
- Oh, S. (2023). Does identity affect labor supply? *American Economic Review*, 113(8):2055–2083.
- Osman, M., McLachlan, S., Fenton, N., Neil, M., Löfstedt, R., and Meder, B. (2020). Learning from behavioural changes that fail. *Trends in Cognitive Sciences*, 24(12):969–980.
- Pandya, S. S. and Venkatesan, R. (2016). French roast: consumer response to international conflict—evidence from supermarket scanner data. *Review of Economics and Statistics*, 98(1):42–56.
- Panzone, L. A., Ulph, A., Hilton, D., Gortemaker, I., and Tajudeen, I. A. (2021a). Sustainable by design: choice architecture and the carbon footprint of grocery shopping. *Journal of Public Policy & Marketing*, 40(4):463–486.
- Panzone, L. A., Ulph, A., Zizzo, D. J., Hilton, D., and Clear, A. (2021b). The impact of environmental recall and carbon taxation on the carbon footprint of supermarket shopping. *Journal of Environmental Economics and Management*, 109:102137.
- Pechey, R., Bateman, P., Cook, B., and Jebb, S. A. (2022a). Impact of increasing the relative availability of meat-free options on food selection: two natural field experiments and an online randomised trial. *International Journal of Behavioral Nutrition and Physical Activity*, 19(1):9.
- Pechey, R., Bateman, P. A., Cook, B., Potter, C., Clark, M., Stewart, C., Piernas, C., and Jebb, S. A. (2022b). Testing the effectiveness of ecolabels to reduce the

- environmental impact of food purchases in worksite cafeterias: A randomised controlled trial. *Appetite*, 179:106277.
- Pechey, R., Hollands, G. J., and Marteau, T. M. (2022c). Explaining the effect on food selection of altering availability: two experimental studies on the role of relative preferences. *BMC Public Health*, 22(1):868.
- Poore, J. and Nemecek, T. (2018). Reducing food’s environmental impacts through producers and consumers. *Science*, 360(6392):987–992.
- Puhani, P. A. (2012). The treatment effect, the cross difference, and the interaction term in nonlinear “difference-in-differences” models. *Economics Letters*, 115(1):85–87.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688.
- Saccardo, S. and Serra-Garcia, M. (2023). Enabling or limiting cognitive flexibility? evidence of demand for moral commitment. *American Economic Review*, 113(2):396–429.
- Sant’Anna, P. H. and Zhao, J. (2020). Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 219(1):101–122.
- Serra-Garcia, M. and Szech, N. (2022). The (in) elasticity of moral ignorance. *Management Science*, 68(7):4815–4834.
- Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, pages 99–118.
- Sparkman, G., Macdonald, B. N., Caldwell, K. D., Kateman, B., and Boese, G. D. (2021). Cut back or give it up? the effectiveness of reduce and eliminate appeals and dynamic norm messaging to curb meat consumption. *Journal of Environmental Psychology*, 75:101592.
- Sparks, P. and Shepherd, R. (1992). Self-identity and the theory of planned behavior: Assessing the role of identification with green consumerism. *Social Psychology Quarterly*, pages 388–399.
- Stigler, G. J. (1961). The economics of information. *Journal of Political Economy*, 69(3):213–225.
- Sunstein, C. R. (2016). People prefer system 2 nudges (kind of). *Duke LJ*, 66:121.
- Sunstein, C. R. (2017). Nudges that fail. *Behavioural Public Policy*, 1(1):4–25.
- Thaler, R. H. (2018). Nudge, not sludge. *Science*, 361(6401):431–431.
- Thaler, R. H. and Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.

- Thaler, R. H., Sunstein, C. R., and Balz, J. P. (2013). Choice architecture. In Shafir, E., editor, *The Behavioral Foundations of Public Policy*, chapter 25, pages 428–439. Princeton University Press, Princeton, NJ.
- Tonke, S. (2025). Shaping identity: Evidence from a large-scale field experiment. *Journal of Political Economy Microeconomics*, 0(ja):null.
- Ursu, R. M. (2018). The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions. *Marketing Science*, 37(4):530–552.
- Weber, E. U. (2013). Doing the right thing willingly. In Shafir, E., editor, *The Behavioral Foundations of Public Policy*, pages 380–397. Princeton University Press, Princeton, NJ.
- Weber, E. U., Johnson, E. J., Milch, K. F., Chang, H., Brodscholl, J. C., and Goldstein, D. G. (2007). Asymmetric discounting in intertemporal choice: A query-theory account. *Psychological Science*, 18(6):516–523.
- Yu, C. W., Zhang, Y. J., and Zuo, S. X. (2021). Multiple switching and data quality in the multiple price list. *Review of Economics and Statistics*, 103(1):136–150.
- Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, 13(1):75–98.

A Appendix: Extra Results

A.1 Summary Statistics and Balance Checks

Table A.1 reports the summary statistics for the sample as a whole as opposed to treatment arm level. For any variables for which datapoints are collected both in week 1 and 2 (e.g. green identity), baseline results from Week 1 are the ones provided.

Table A.1: Overall Week 1 sample descriptive statistics

	Summary
	875
Female	
No	439 (50.2%)
Yes	436 (49.8%)
Age group	
Age <=31	202 (23.1%)
Age 32-41	283 (32.3%)
Age 42-52	185 (21.1%)
Age 53+	205 (23.4%)
University educated	
No	297 (33.9%)
Yes	578 (66.1%)
Income group	
Income <GBP30k	405 (46.3%)
Income GBP30-50k	300 (34.3%)
Income GBP50k+	170 (19.4%)
No children	
No	343 (39.2%)
Yes	532 (60.8%)
Shops for others	
No	175 (20.0%)
Yes	700 (80.0%)
No dietary restrictions	
No	50 (5.7%)
Yes	825 (94.3%)
Hunger category	
High hunger	375 (45.5%)
Medium hunger	154 (18.7%)
Low hunger	296 (35.9%)
Basket total GHG (kg CO2e)	7.29 6.56 (3.95)
Checkout spend (GBP)	39.16 40.54 (8.68)
Items bought	18.83 18.00 (6.07)
Search effort (minutes)	11.24 8.83 (14.37)
Consumer satisfaction	6.61 7.00 (2.22)

Overall Week 1 sample descriptive statistics.

Categorical variables report frequency (%) by category level.

Continuous variables report mean, median and SD (in parentheses).

Table A.2 reports a longer set of balance checks across variables collected in Week 1. They are presented by treatment arm and include tests of significant differences across arms. The balance checks indicate no significant differences across arms for

any of the variables collected in Week 1.

Table A.2: Descriptive statistics and randomisation balance checks (Week 1)

	Control n=295 (33.7%)	Pre-MPL Cue n=296 (33.8%)	Post-MPL Cue n=284 (32.5%)	Test
Female	146 (49.5%)	150 (50.7%)	140 (49.3%)	0.937
Age: less than 31 y.o.	61 (20.7%)	68 (23.0%)	73 (25.7%)	0.627
University Educated	200 (67.8%)	192 (64.9%)	186 (65.5%)	0.731
Income Group: less than £30k	138 (46.8%)	122 (41.2%)	145 (51.1%)	0.188
No Children	177 (60.0%)	181 (61.1%)	174 (61.3%)	0.942
Shops for Others	243 (82.4%)	227 (76.7%)	230 (81.0%)	0.198
No Dietary Restrictions	278 (94.2%)	281 (94.9%)	266 (93.7%)	0.804
Hunger Category	124 (44.9%)	124 (44.1%)	127 (47.4%)	0.825
Impatient	107 (36.3%)	124 (41.9%)	121 (42.6%)	0.231
Risk Averse	143 (48.5%)	145 (49.0%)	143 (50.4%)	0.897
Low CRT	127 (43.1%)	102 (34.5%)	116 (40.8%)	0.085
Internal Locus	134 (45.4%)	149 (50.3%)	133 (46.8%)	0.469
Low Self-Control	120 (40.7%)	120 (40.5%)	118 (41.5%)	0.965
Low Green Self Wk1	150 (50.7%)	152 (51.4%)	132 (46.5%)	0.458
Low Green Norms Wk1	145 (49.0%)	154 (52.0%)	138 (48.6%)	0.672
Low Green Efficacy	171 (58.0%)	156 (52.7%)	156 (54.9%)	0.434
Low Paternal Need	121 (41.0%)	127 (42.9%)	104 (36.6%)	0.287
Government Control Aversion	118 (40.0%)	130 (43.9%)	99 (34.9%)	0.082
Basket Total GHG (kg co2e)	7.41 (3.99)	7.31 (4.16)	7.14 (3.69)	0.704
Search Effort (mins)	10.77 (7.23)	11.93 (21.51)	11.02 (10.01)	0.585

Joint significance balance tests across all three treatment arms are reported.

Categorical variables are reported with number of participants and in parentheses the share in category = 1 (for Hunger Category, category = Low, below median). P-value from Pearson chi-squared test.

2 category variables: Female, University Educated, No Children, Shops for Others, No Dietary Restrictions, Impatient, Risk Averse, Low CRT, Internal Locus, Low Self-Control, Low Green Self Wk1, Low Green Norms Wk1, Low Green Efficacy, Low Paternal Need, Government Control Aversion. 3 category variables: Income Group, Hunger Category. 4 category variables: Age Group. For continuous variables: Mean (SD): p-value from linear regression.

Table A.3 and Table A.4 compare key socio-demographic, preference and shopping behaviour characteristics between the Seekers and Avoiders of the eco-sorting layout (see Section 6.1, those with 1 or no changes in their WTP Multiple Price List). Table A.3 compares the Seekers and Avoiders across that whole sample, while Table A.4 focuses on a subsample of the more marginal participants, those with a WTP between -1 and 1, which are those most likely targeted with the identity cue.

Table A.3: Comparison of Avoiders (WTP<0) and Seekers (WTP>=0)

	Seekers (WTP>=0) n=299 (54.8%)	Avoiders (WTP<0) n=247 (45.2%)	p-value
Age group			
Age <=31	67 (22.4%)	58 (23.5%)	0.480
Age 32–41	93 (31.1%)	81 (32.8%)	
Age 42–52	58 (19.4%)	55 (22.3%)	
Age 53+	81 (27.1%)	53 (21.5%)	
Female	0.51 (0.03)	0.49 (0.03)	0.668
University educated	0.71 (0.03)	0.61 (0.03)	0.013
No children	0.61 (0.03)	0.60 (0.03)	0.809
Income group			
Income <GBP30k	147 (49.2%)	116 (47.0%)	0.804
Income GBP30–50k	95 (31.8%)	85 (34.4%)	
Income GBP50k+	57 (19.1%)	46 (18.6%)	
Green self score	21.34 (0.36)	16.79 (0.45)	<0.001
Basket CO2e (kg, Wk1)	7.02 (0.22)	7.63 (0.25)	0.066
Checkout spend (GBP, Wk1)	39.38 (0.51)	39.69 (0.55)	0.687
Items in basket (Wk1)	19.41 (0.36)	18.79 (0.39)	0.243
Search effort (min, Wk1)	11.43 (0.47)	10.40 (0.62)	0.175
Shops for others	0.83 (0.02)	0.81 (0.03)	0.412
No dietary restrictions	0.92 (0.02)	0.96 (0.01)	0.056

Sample: single-switchers and non-switchers (n_changes <= 1).

Mean (SE): p-value from linear regression on avoider dummy.

Frequency (%): p-value from Pearson chi-squared test.

Table A.4: Comparison of Avoiders (WTP<0) and Seekers (WTP>=0), |WTP| <= 1.0

	Seekers (WTP>=0) n=205 (51.4%)	Avoiders (WTP<0) n=194 (48.6%)	p-value
Age group			
Age <=31	46 (22.4%)	48 (24.7%)	0.256
Age 32–41	64 (31.2%)	64 (33.0%)	
Age 42–52	39 (19.0%)	45 (23.2%)	
Age 53+	56 (27.3%)	37 (19.1%)	
Female	0.49 (0.04)	0.47 (0.04)	0.713
University educated	0.69 (0.03)	0.62 (0.03)	0.147
No children	0.62 (0.03)	0.61 (0.04)	0.822
Income group			
Income <GBP30k	101 (49.3%)	90 (46.4%)	0.771
Income GBP30–50k	66 (32.2%)	69 (35.6%)	
Income GBP50k+	38 (18.5%)	35 (18.0%)	
Green self score	20.44 (0.45)	16.91 (0.51)	<0.001
Basket CO2e (kg, Wk1)	7.00 (0.26)	7.65 (0.29)	0.093
Checkout spend (GBP, Wk1)	39.22 (0.62)	39.37 (0.64)	0.865
Items in basket (Wk1)	19.27 (0.44)	18.73 (0.45)	0.394
Search effort (min, Wk1)	11.74 (0.61)	10.42 (0.76)	0.174
Shops for others	0.84 (0.03)	0.78 (0.03)	0.121
No dietary restrictions	0.94 (0.02)	0.96 (0.01)	0.323

Sample: single-switchers and non-switchers (n_changes <= 1), WTP in [-1.0, 1.0].

Mean (SE): p-value from linear regression on avoider dummy.

Frequency (%): p-value from Pearson chi-squared test.

A.2 Effects of Cue on WTP

Figure A.1 shows the distribution of WTP for the eco-sorting layout across the identity-cue treatment arms. Notably, we distinguish between groups that received the cue before the WTP elicitation (Pre-MPL Cue) and those that received it after or not at all (No Pre-MPL Cue). The sample is restricted to participants with at most three preference reversals in the WTP Multiple Price List (MPL). We observe a clear shift in the WTP distribution for the Pre-MPL Cue group, which is statistically significant as presented in the results section with the conditional regressions estimates.

Table A.5 reports the same regression estimates as in Table 3 but with the full list

of control variable estimates, which are also discussed in the results section.

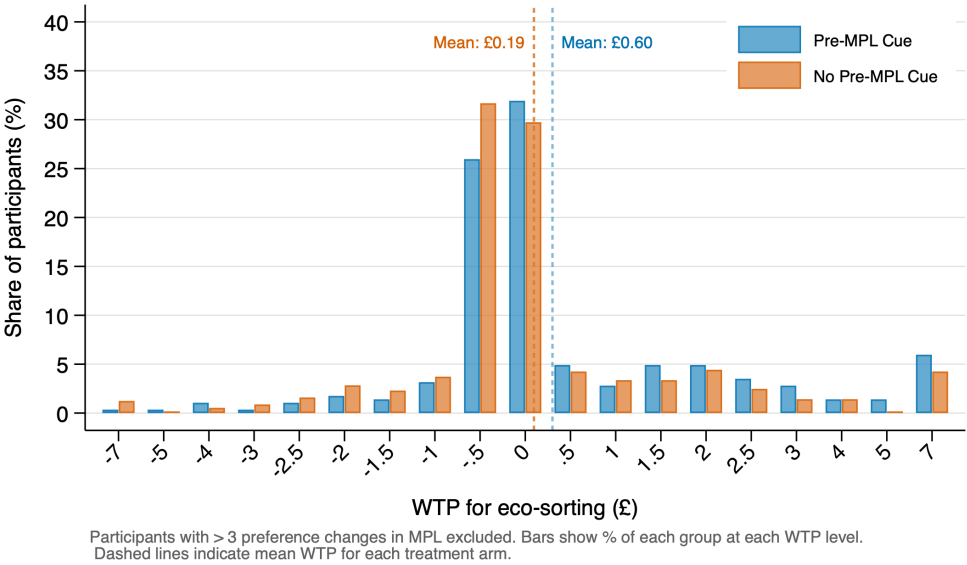


Figure A.1: Distribution of WTP for the eco-sorting layout by treatment group

Table A.5: Effect of identity cue on WTP for eco-sorting, with control estimates

	Simple	Socio controls	Full controls	Full controls w/ Multiple-switchers
Pre-MPL cue	0.404*** (0.156)	0.416*** (0.155)	0.453*** (0.154)	0.457*** (0.152)
Female		0.225 (0.149)	0.122 (0.148)	0.110 (0.146)
University educated		0.281* (0.170)	0.223 (0.167)	0.215 (0.164)
No children		0.033 (0.158)	0.074 (0.170)	0.082 (0.167)
Income £30–50k		-0.028 (0.172)	0.063 (0.165)	0.064 (0.163)
Income £50k+		-0.207 (0.177)	-0.034 (0.175)	-0.017 (0.173)
Age 32–41		0.041 (0.188)	-0.067 (0.185)	-0.059 (0.181)
Age 42–52		0.221 (0.211)	0.046 (0.204)	0.036 (0.201)
Age 53+		0.300 (0.237)	0.159 (0.237)	0.133 (0.231)
Low green self-identity			-0.607*** (0.143)	-0.609*** (0.140)
Low search effort			-0.175 (0.150)	-0.171 (0.147)
No dietary restrictions			-0.751* (0.442)	-0.648 (0.430)
Shops for others			0.014 (0.213)	-0.027 (0.208)
Items in basket			0.048*** (0.017)	0.048*** (0.017)
Checkout spend (£)			-0.021** (0.010)	-0.021** (0.010)
Participants	827	827	827	849

Columns 1-3 include only those with 1 or 0 preference reversals in the MPL.

Column 4 uses the fuller sample of those with 3 or fewer preference reversals.

Reference categories: age ≤ 31 , income $< £30,000$.

Green self-identity and search effort (both relative to median of population) are measured in week 1.

Items bought and checkout spend are from week 1 as well.

OLS with cluster-robust standard errors (id-level) in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.2.1 Inverted-U effect of Cue on WTP

The table and figure in this subsection explore the possibility of an inverted-u shape in the effect of the identity cue on WTP relative to participants' baseline green identity, as suggested by the Bénabou and Tirole (2011) identity model. The results of these are discussed in Section 6.2.

Figure A.2 below shows some indications of a hump-shaped pattern. The identity cue's effect is significantly positive across the central portion of the distribution and loses significance at the tails, but we also note that the tails are less populated and have wider confidence intervals.

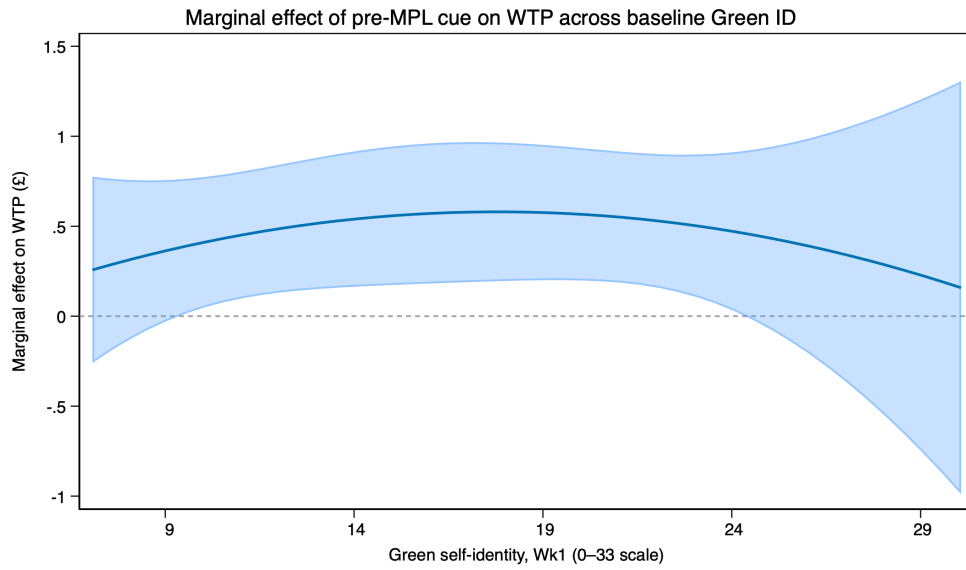


Figure A.2: Inverted-U effect of Cue on WTP by baseline green identity

Notes: Marginal effect of the pre-MPL cue on WTP (£) across baseline green self-identity (week 1, 3-33 scale), from the quadratic specification in Table A.6, Col. 5 ($n = 828$, $n_{\text{changes}} \leq 1$). Shaded band: 95% CI, cluster-robust SEs (id-level).

Table A.6: Heterogeneous effect of the pre-MPL cue on WTP by baseline green self-identity

	Tercile CATE		Quadratic (inverted-U)	
	≤ 1	≤ 3	≤ 1	≤ 3
Pre-MPL cue	0.499*** (0.184)	0.499*** (0.183)	0.576*** (0.192)	0.569*** (0.188)
Mid Green ID (Terc. 2)	0.452** (0.182)	0.473*** (0.180)		
High Green ID (Terc. 3)	1.126*** (0.224)	1.064*** (0.215)		
Cue $\times$ Mid Green ID	0.145 (0.334)	0.126 (0.329)		
Cue $\times$ High Green ID	-0.389 (0.400)	-0.332 (0.396)		
Green ID (centred)			0.067*** (0.014)	0.064*** (0.014)
Cue $\times$ Green ID			-0.007 (0.028)	-0.003 (0.027)
Green ID ²			0.003* (0.002)	0.003* (0.002)
Cue $\times$ Green ID ²			-0.003 (0.003)	-0.003 (0.003)
Socio-demographic controls	Yes	Yes	Yes	Yes
Shopping controls	Yes	Yes	Yes	Yes
Observations	827	849	827	849

All columns use the full-controls specification of the main regression table in the text.

Columns ≤ 1 and ≤ 3 use samples with at most 1 or 3 MPL preference reversals, respectively.

The lowest Wk1 green identity tercile is the reference group.

Green ID is mean-centred for the quadratic specifications.

OLS with cluster-robust standard errors (id-level) in parentheses.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.3 Mechanism Check: Green-Identity Updating

A.3.1 Cue on Green Identity Updating

We now consider the mechanism through which the identity cue raises WTP for the eco-sorting layout. Under the Bénabou and Tirole (2011) identity model, the cue should be working through participants' self-image, nudging their green identity upward, which then feeds back into preferences and behaviour (Bénabou and Tirole, 2011; Tonke, 2025). To check that the WTP effect we report in Section 6.2 is indeed identity-driven, and not picked up through some other channel, we look at the change in self-reported green identity from week 1 to week 2 across the two cue arms (pre-MPL and post-MPL) and the no-cue control.

With Table A.3.1 we estimate the effect of the identity cue on green self-identity using a panel fixed-effects difference-in-differences (FE DiD) specification, with individual fixed effects, person-week observations and standard errors clustered at the participant level. The dependent variable is self-reported green identity on the 3–33 scale.

Column 1 of Table A.3.1 pools the two cue arms into a single “any cue” indicator and compares them to the no-cue control. Pooled, having received an identity cue raises green identity by 0.83 points on the 3–33 scale (at the 1%-level), against a baseline week 1 control-arm mean of 19.0 points, and so the pooled cue effect represents around a 4% increase on baseline. Columns 2 and 3 report each cue arm separately against the no-cue control, dropping the participants in the other cue arm. Only the pre-MPL cue is associated with a statistically significant identity update, with a point estimate increase of 1.11 (at the 1%-level) in column 2; the post-MPL cue point estimate in column 3 is positive but smaller (0.55) and not significant. Column 4 enters both cue arms simultaneously on the full sample, which delivers virtually the same coefficients (1.11 and 0.55). Column 5 re-estimates

the joint cue specification of column 4 on the IPW sample we use for the basket-emissions analysis (single-switchers and excluding those that selected always eco or standard layout in the MPL (so layout assignment was not randomised, stochastic), reweighted by the design MPL draw probabilities of Section 5.2) to equalise the layout-assignment probability across the cue arms, and subsequently any effect on green identity beliefs in week 2. The pre-MPL cue coefficient is essentially unchanged (1.17, $p < 0.01$) and the post-MPL cue coefficient now also reaches the 5% level (0.74). The gain in statistical significance is driven by the sample change for the IPW, rather than the re-weighting itself, as shown in Table A.3.1. Column 2 and 3 use the same sample (the IPW sample) but column 2 does not apply the IPW reweighting, so only presents the estimates for the change in sample, while column 3 uses both the IPW sample and reweighting.

The WTP effect of the identity cue documented in Section 6.2 shows a significant upward shift in self-reported green identity, which is in support with the identity-updating mechanism in Bénabou and Tirole (2011) and Tonke (2025) where an identity-cue was also used to update identity, but the identity-updating mechanism could not be tested directly in that setting.

We also note the possibility, but no statistical evidence of a timing effect between the two cue treatment-arms. This should be for future exploration. This suggestion is based on the identity update being concentrated in the pre-MPL cue arm (1.11–1.17), while the post-MPL cue produces a smaller shift (0.55–0.74) which only crosses the 5%-level once we move to the IPW sample in column 5. The formal pre-vs-post timing test from the IPW joint specification in column 5 is $\hat{\beta}_{\text{pre}} - \hat{\beta}_{\text{post}} = 0.43$ (Wald Test, $\text{SE} = 0.38$, $p = 0.25$), so we cannot statistically separate the two cue coefficients with the precision available here. We note however, the experiment was not powered for this question.

A final point worth flagging is that the week 2 identity is reported after participants

have been through the MPL and the eco- or standard-sorted shop, so in principle the cue's effect on identity could be affected indirectly through the sorting layout assigned rather than directly. We think this is not what is happening here. The IPW specification in column 5 reweights the sample to equalise the layout-assignment probability across the cue arms, breaking the cue–layout link, and yet the pre-MPL cue coefficient barely moves (from 1.11 in column 4 to 1.17 in column 5). If sorting exposure were mediating the identity update, reweighting away that channel should shrink the coefficient; it does not. We take this as evidence that the cue's effect on green identity is direct, with the layout assignment that follows the cue not doing the work.

Table A.7: Effect of Identity Cue on Green Self-Identity

	Any cue	Pre-cue	Post-cue	Joint	Joint (IPW-ATE)
Any cue × Post	0.831*** (0.312)				
Pre-MPL cue × Post		1.105*** (0.364)		1.105*** (0.364)	1.179*** (0.390)
Post-MPL cue × Post			0.546 (0.356)	0.546 (0.356)	0.748** (0.374)
Observations	1750	1182	1158	1750	1556
Participants	875	591	579	875	778

FE DiD with individual fixed effects and clustered standard errors (in parentheses).

Green self-identity (3–33 scale). Observations are person-weeks (Post = Week 2).

Col. 1: Any cue (pre or post arm pooled) vs no-cue control.

Col. 2-3: Each cue arm vs the no-cue control only (other arm is excluded from sample).

Col. 4: Full sample with both cue arms entered simultaneously.

Col. 5: Same joint cue model as Col. 4, estimated on the IPW-trimmed eco-sorting sample.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.8: Decomposing the IPW Robustness Check: Sample Restriction vs. Reweighting

	Full sample (Col. 4 of A.7)	IPW sample, unweighted	IPW sample, ATE-weighted (Col. 5)
Pre-MPL cue $\times$ Post	1.105*** (0.364)	1.213*** (0.388)	1.179*** (0.390)
Post-MPL cue $\times$ Post	0.546 (0.356)	0.769** (0.374)	0.748** (0.374)
Observations	1750	1556	1556
Participants	875	778	778

FE DiD with individual fixed effects and clustered standard errors (in parentheses).

All columns estimate the joint cue specification from Col. 4 of Table A.7.

Col. 1 reproduces Table A.7 Col. 4 (full sample, unweighted).

Col. 2 restricts to the IPW-ATE sample (single-switchers, excl. corner non-switchers: $\hat{p}(w_i) \in (0, 1)$), but does not apply the IPW weights.

Col. 3 reproduces Table A.7 Col. 5 (same sample as Col. 2, ATE-reweighted)

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A.3.2 Eco-Sorting on Green Identity Updating

As discussed in section 6.4, the eco-sorting layout could also have an effect on green identity updating, if being assigned to the eco-sorted layout and shopping greener, consciously, as a result reinforces the cue's identity message (Bénabou and Tirole, 2011).

We test this in the same FE DiD specification used for the cue-only mechanism check, but now adding the eco-sorting indicator and its interaction with the post-treatment period. We restrict all our regressions to the IPW sample for sample consistency. Column 1 and 2 of Table A.9 present the main effects of the eco-sorting layout with and without the IPW reweighting, while column 3 only uses the control (no-cue) arm, to test the effect of the eco-sorting layout, still using the IPW weights, and as a sub-sample of the IPW sample. Finally, column 4 adds the three-way interaction between the eco-sorting layout, the post-treatment period and the

cue indicator.

Across all four specifications, we find no evidence of the eco-sorting layout having an effect on green identity updating, while it did have reported effects on behaviour (change in basket emissions). This would seem to suggest that by acting on behaviour through unconscious decision processes, the eco sorting layout does not reinforce the green-self image, nor does it interact with the cue to amplify identity for the cohort that received the cue. We note however that columns 3 and 4 are underpowered. The no-cue arm in column 3 (is only $n = 254$) and the triple interaction in column 4 would be sample size demanding, so we cannot draw strong conclusions from those inferences.

Table A.9: Effect of Eco-Sorting Layout on Green Self-Identity

	Eco-sorting (IPW-ATE)	Eco-sorting (IPW sample, no weights)	Eco-sorting (no-cue arm, IPW-ATE)	Cue $\times$ Eco-sorting (IPW-ATE)
Eco-sorting $\times$ Post	-0.158 (0.312)	-0.060 (0.312)	-0.104 (0.544)	-0.104 (0.543)
Any cue $\times$ Post				0.951** (0.462)
Cue $\times$ Eco-sorting $\times$ Post				0.013 (0.661)
Observations	1556	1556	508	1556
Participants	778	778	254	778

FE DiD with individual fixed effects and clustered standard errors (in parentheses).

Outcome: Green self-identity (3–33 scale). Observations are person-weeks (Post = Week 2).

(Cols. 1, 2, 4): IPW analysis sample (single-switchers, excl. corner non-switchers, $\hat{p}(w_i) \in (0, 1)$).

Col. 1: Eco-sorting effect, ATE-reweighted ($1/\hat{p}$ for treated, $1/(1 - \hat{p})$ for control).

Col. 2: Same sample, unweighted — isolates the role of the IPW reweighting.

Col. 3: No-cue arm only, ATE-reweighted: layout-only effect free of any cue-driven identity update.

Col. 4: Three-way Cue $\times$ Eco-sorting $\times$ Post interaction on the IPW sample:

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

A.4 Robustness Checks: Eco-sorting layout on emissions

A.4.1 CATE by wtp type robustness

Table A.10 and Table A.11 verify the results from Table 5 in the main text results on the impact of the eco-sorting layout on basket emissions, Section 6.3.

Table A.10 tests the same CATE categories (Avoiders, Indifferent, Seekers) but with the fuller sample ($n_{\text{changes}} \leq 3$) and without the IPW weighting. We observe that the point estimates and their significance are very consistent with those in Table 5, with the exception that now the ‘Seekers’ group’s estimate is also significant at the 1% level, which is perhaps driven by the larger sample size, while the estimated coefficient remains very similar.

Table A.11 follows suggestions by Sant’Anna and Zhao (2020) by using their doubly-robust regression adjustment to the IPW DiD specification. The Sant’Anna and Zhao (2020) doubly-robust method for difference-in-differences calculates the regression estimates by adjusting them with both the propensity score weights (IPW) (making treated and control populations comparable) and an outcome regression model that measures any heterogeneity in outcome conditional on participant characteristics that can then be applied as predictions to those controls in the treated groups (providing information on heterogeneity in trends, rather than assuming an average parallel trend across the sample). The outcome regression model estimates the expected outcome in the control group over time conditional on supplied controls (e.g. socio-demographics, and green preferences pre-treatment in our case). This method combines the IPW reweighting with the corrected outcome regression trend prediction adjustments, making the estimates “doubly robust”. Again, we observe robust estimates for all three groups (compared to our main Table 5 and the other robustness Table A.10). As with the same smaller sample (required for the IPW weights), the ‘Seekers’ group estimate is stable but

the inference is seemingly weakened by the decreased group sample size.

Table A.10: Robustness: CATE on wider sample ($n_changes \leq 3$, unweighted)

	Avoiders	Indifferent	Seekers
Eco-sorting $\times$ Post	-0.262*** (0.065)	-0.461*** (0.074)	-0.207*** (0.076)
Observations	710	518	470
% change (Puhani)	-23.1	-37.0	-18.7
Participants	355	259	235

Two-way fixed-effects DiD on log basket CO2e. No IPW weights.

Sample: $n_changes \leq 3$, WTP in $[-7, 7]$ (corner switchers included).

Cluster-robust SEs at participant level.

Joint test of heterogeneity across WTP types: $p = 0.037$.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A.11: Doubly-robust (AIPW) DiD by WTP type, headline IPW sample

	Avoiders	Indifferent	Seekers
Eco-sorting $\times$ Post	-0.270*** (0.067)	-0.447*** (0.081)	-0.160* (0.090)
Observations	678	514	364
% change (Puhani)	-23.6	-36.0	-14.8
Participants	339	257	182

Sant'Anna & Zhao (2020) doubly-robust DiD (ATT estimand).

Restricts to the IPW-ATE sample (single-switchers,
excl. corner non-switchers: $\hat{p}(w_i) \in (0, 1)$)

Covariates: socio-demographics (gender, education, children,
income group, age group), and pre-treatment (Wk1) green self-identity.

Cluster-robust SEs at participant level.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

B Appendix: Managerial and Policy Implications

B.1 Negative Spillovers

In this Appendix section, we include any extra tables or figures relevant for the managerial and policy implications discussion in Section 7 of the main text.

Table A.12: Effect of eco-sorting on checkout spend and search effort: IPW FE DiD

	Checkout spend (£)		Search effort (minutes)	
	Baseline	+ Cue control	Baseline	+ Cue control
Eco-sorting $\times$ Post	-2.504*** (0.452)	-2.504*** (0.452)	-0.417 (1.011)	-0.492 (1.064)
Cue $\times$ Post		-0.000 (0.506)		-1.191 (1.132)
Items bought	0.896*** (0.074)	0.896*** (0.074)	0.250** (0.114)	0.248** (0.115)
Observations	1556	1556	1556	1556
Participants	778	778	778	778

Two-way fixed-effects (week and individual) DiD on the IPW sample

(single-switchers, $\hat{p}(w_i) \in (0, 1)$), reweighted with IPW weights.

Cluster-robust SEs at the participant level (in parenthesis).

Items bought retained as a time-varying control.

+Cue control columns add an interaction of cue exposure (pre or post)

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

B.2 Heterogeneity in Effects

In this subsection, we include the figures produced to assess for heterogeneity of effects from the eco-sorting across age, gender, income, education parameters.

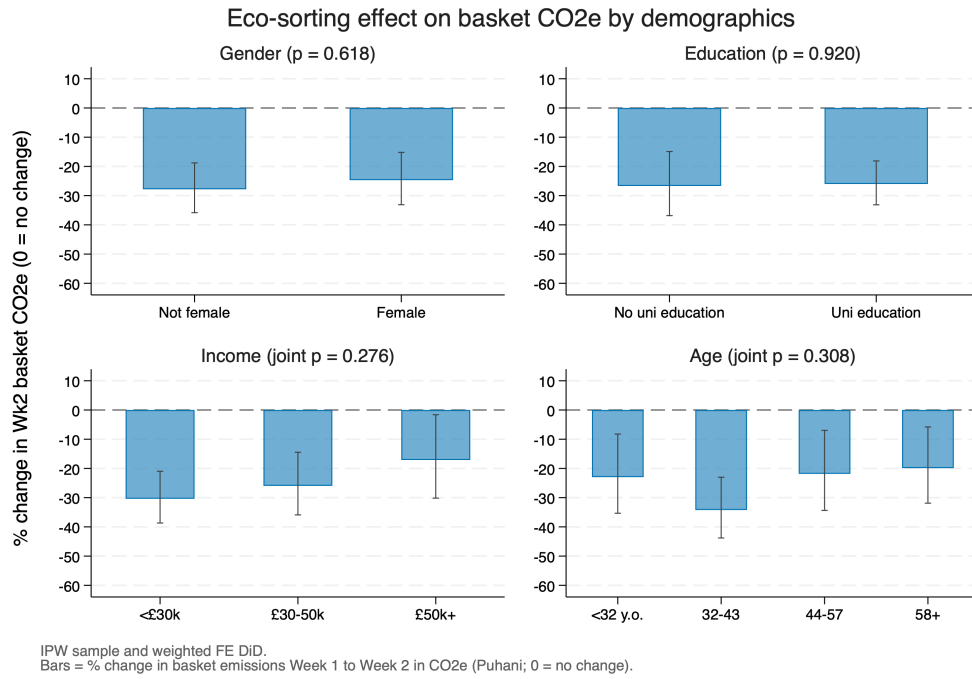


Figure A.3: Eco-sorting effect by demographics

Note: ATE in percentage change in basket emissions, computed with Puhani's conversion as $(\exp(\delta) - 1) \times 100$, where δ are the IPW (sample and weightings) FE-DiD log coefficients on Eco-sorting.

B.3 Aggregate emissions savings under different deployment scenarios

Figure A.4 provides two complementary perspectives on the aggregate emissions implications of the intervention combination. Panel (a) contextualises the scale of the eco-sorting layout's effect relative to the UK's dietary emissions targets. The current sample mean of 7.29 kgCO₂_e per weekly basket falls to 5.36 kg under the eco-sorted layout, a reduction of 26.5%. The dashed lines show the CCC 2050 target of 1.82 kgCO₂_e per weekly basket implied by the Sixth Carbon Budget, illustrating both the scale of ambition required and the direction in which the intervention moves behaviour. Panel (b) translates the causal effect into expected emissions savings per shopper per week across four deployment scenarios, each varying by the

share of consumers who adopt the eco-sorted layout. The calculation in each case is: uptake rate $\times$ 26.5% $\times$ 7.29 kgCO_{2e}. The scenarios are designed to reflect realistic deployment contexts, analogous to the choice a consumer faces when setting up an online account where the layout preference is set once and persists unless actively changed.

The four scenarios are as follows. *Passive opt-in* (approximately 5% uptake) represents a lower bound in which consumers must actively seek out the eco-sorted layout with no prompt, yielding an expected saving of 0.10 kgCO_{2e} per shopper per week. *No cue* (54% uptake) uses the observed share of participants in our experiment with WTP at or above zero, that is, those who are indifferent or actively prefer the eco-sorted layout, and represents a prompted opt-in scenario in which the layout is offered at no cost. This yields an expected saving of 1.06 kgCO_{2e} per shopper per week. *With cue* (64% uptake) adds the identity cue, which shifts approximately 10 percentage points of participants across the opt-in threshold, raising expected savings to 1.24 kgCO_{2e} per shopper per week, a gain of 17% relative to the no-cue scenario. *Default opt-out* (approximately 95% uptake) applies the eco-sorted layout universally unless consumers actively switch back, yielding the maximum expected saving of 1.83 kgCO_{2e} per shopper per week.

The figure illustrates the complementarity argument directly. Adding the identity cue to a voluntary opt-in scenario generates a 17% gain in aggregate emissions savings at no additional cost to consumer autonomy. While the default opt-out scenario achieves the largest absolute reduction, it does so by applying the layout to a meaningful share of participants who would prefer not to have it. The cue narrows this gap substantially while preserving the voluntary character of the adoption decision.

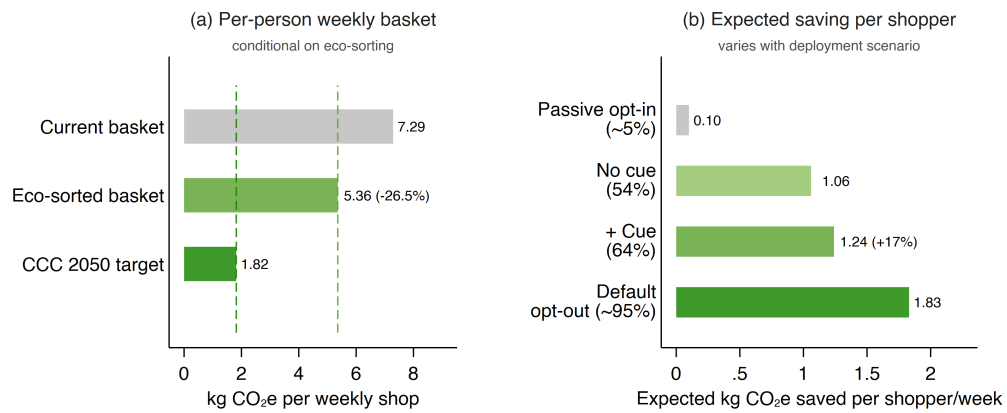


Figure A.4: Eco-sorting effect in context

C Food and Basket Variables

C.1 Participant Basket Results

Week 1 - Grocery Basket Distribution Participant's grocery baskets in week 1 are right-skewed in terms of their CO₂e emissions, with an average of 7.29 kg CO₂e and a median of 6.56 kg CO₂e. The skew test (D'Agostino-Pearson test) and Shapiro-Wilk test both reject the null hypothesis of normality for the distribution of week 1 basket emissions at the 1% significance level, as indicated in the footnote of A.5.

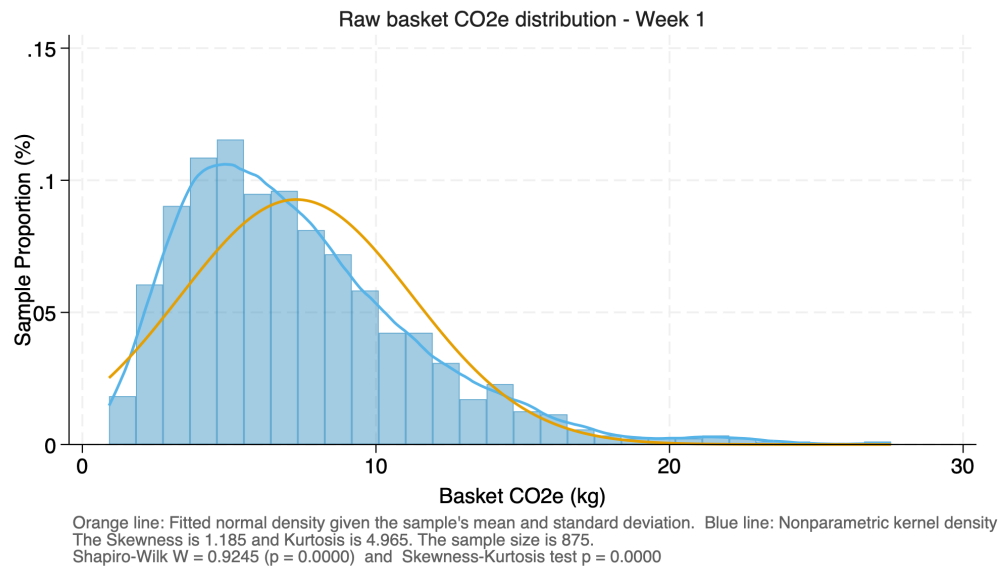


Figure A.5: Testing Normality of Raw basket CO₂e distribution – Week 1

Grocery basket change within-person distribution With Figure A.6 we compare the distribution of the raw change in basket emissions per person (week 2 - week 1) and the log of the same change in basket emissions. Both distributions are still skewed but the log transformation reduces the skewness and the tail of the distribution.

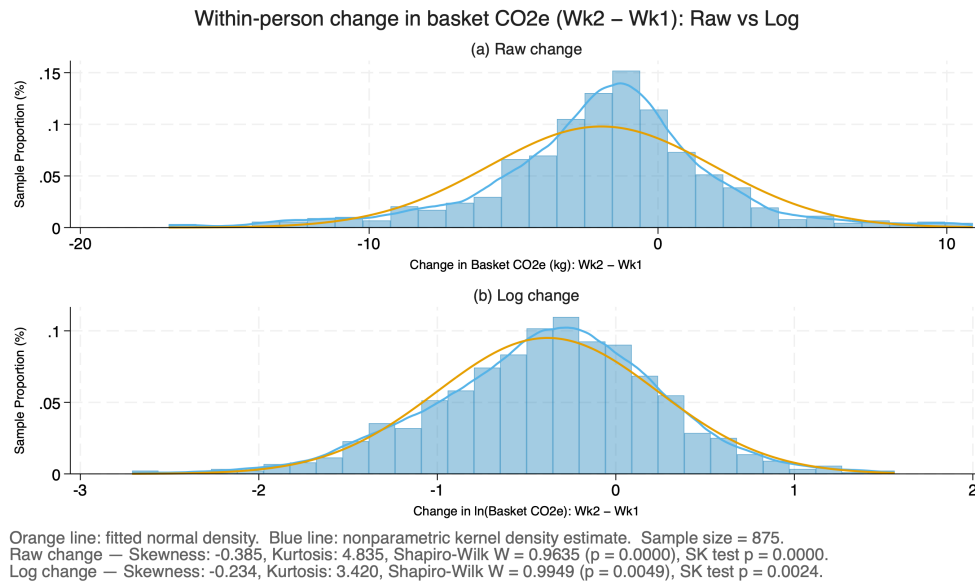


Figure A.6: Tail and Normality of Within-Person Distributions: Raw vs Log

D Appendix: Experiment Visuals

D.1 Identity Cue Message

The Figure A.7 pop-up was only shown to participants in the ‘cue’ treatments. For those participants it was shown at the beginning of the week 2 experiment, before accessing the shopping environment. Depending on the sub-treatment group, it was either shown before or after the willingness to pay for the eco and non-eco sorting environment multiple price list exercise.



Figure A.7: Image of Identity Cue

D.2 Shopping Environment

These two visuals are screenshots taken of the Woods Experimental online shopping platform. One provides a view of the shopping list appended to the side of the website portal, with an example of search results after searching for 'pizza' in the standard layout. The other screenshot highlights the filtering options also available with the Woods platform. In both cases, the food labels and prices are also visible, as were shown to the participants.

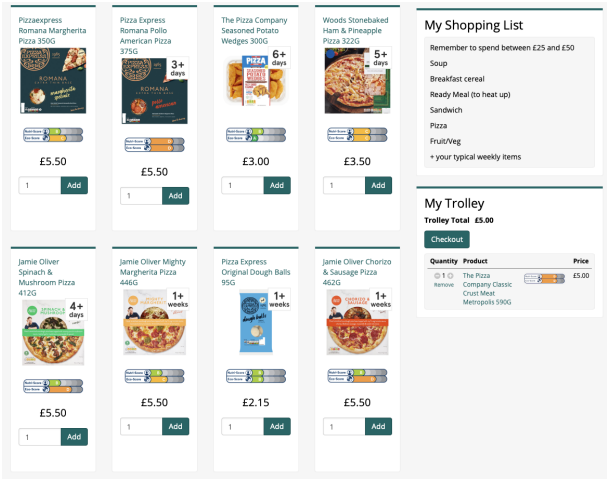


Figure A.8: Image of shopping list provided

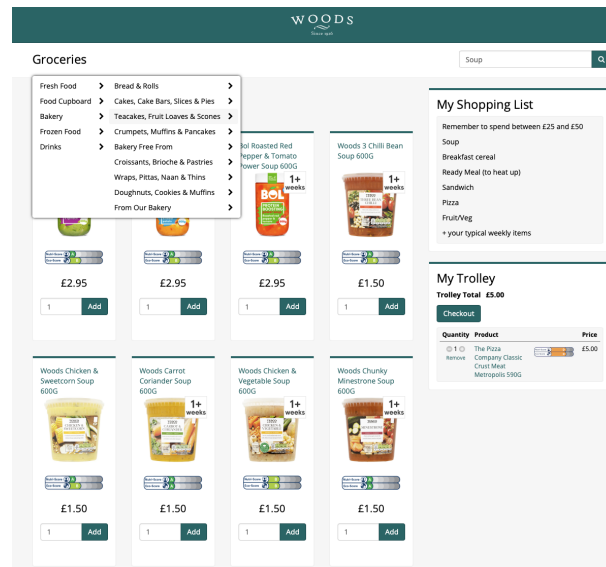


Figure A.9: Image of Shopping Environment with filters

D.3 Week 1 Study Description: Inviting Participation

Prolific Project Title: *Online Shopping - Research Project*

In this online research project (that is in two stages over 10 days), we will ask you to answer several survey questions as well as do some online grocery shopping on our online experimental shopping platform.

Each week, you will start with a survey, then be sent to our online experimental shopping platform, then we will collect a few more answers from you on a follow-up survey to which you will be sent from the shopping platform. This will be repeated in a second week. All information is completely anonymous, and no questions will make you identifiable.

You can complete these surveys and shopping tasks at any point, within the time-frame, that is convenient to you.

Each week, both surveys and the shopping task together will take roughly 18 minutes.

We request that if you participate, you use a tablet or computer.

Each week completed you will be paid £2 for your time, with a bonus of £1 if you complete both weeks. IN ADDITION, you will be entered to a random draw for a chance to win the contents of one of your shopping baskets, and any unspent shopping budget allocated to you, for real.

As an overview, of the project timeline:

Week 1: Survey + Shopping Basket + Survey: roughly 18 mins: £2.

Week 2: Survey + Shopping Basket + Survey: roughly 18 mins: £2 + £1 bonus for returning.

Please, only consider this project if you think it is highly likely that you will be available both weeks.

The researchers are very grateful for your time and consideration.

We recommend that you only take this study with the use of a tablet, laptop or desktop computer.

Unfortunately, at the moment, we do not have sufficient Vegan, Halal or Kosher grocery options, so if these are relevant for your diet, we do ask that you do not take this study.

Thank you.

D.4 Week 1 Grocery Shopping Task for Participants

In this next section, you will complete a weekly food shop on our online platform. You will have a virtual budget of £50 to spend on items you typically purchase. It is important that you make true choices, as there is a chance that you will actually receive the order you place.

You must spend between £25 and £50 on items in the Woods platform. You have a chance to win the items for real, and any unspent budget (£50) will be transferred to you in cash.

Please purchase groceries that you would typically purchase in your weekly shop. In your shopping basket you must also include all the following: Soup - Breakfast Cereal - Ready Meal (to heat up) - Sandwich - Pizza - Fruit/Veg. You can browse the platform just like in real life: view product details, add or remove items, and adjust your basket as needed. These items will also count toward your budget.

Shopping Prize Draw: A random draw will determine 5 winners. Winners will be notified via Prolific within a few weeks after the study ends. If selected, your shopping basket will be prepared for collection at your local Tesco store. Your remaining budget will also be transferred to you.

When you are ready, click to the next page and you will be brought to the supermarket task.

After you have finished the supermarket task, you will be brought back to this survey for some follow-up questions and the completion code.

E Multiple Price List and Willingness-to-pay Variable Construction

This appendix describes the Multiple Price List (MPL) elicitation task and the construction of the Willingness-to-pay (WTP) measure. Section E.1 presents the MPL design and comprehension check. Section E.2 describes WTP construction procedure. Section E.4 details the cleaning rules and treatment of remaining observations.

E.1 Multiple Price List (MPL) Description

Figure A.10 reproduces the screenshots shown to participants in week 2, including the comprehension check. Panels A.10a and A.10b show the MPL introduction and the incentive instructions. Panel A.10c provides an illustrative example of the MPL format (presented prior to the study-specific context). Panel A.10d shows the comprehension check.³⁴

Participants who failed the comprehension were notified and required to retake the check; they could not proceed to the incentivised MPL without passing. We record the number of attempts required to pass. In our final sample, 93% of participants passed on the first attempt (814 of 875). Of the 62 participants who required a second attempt, 10 (1.1% of the final sample) failed again and required a third attempt.

The incentivised MPL used to elicit preferences over the shopping sorting options is shown in Figure A.11.

³⁴The comprehension test asked “If you choose Option B (the right-hand option) in the example decision below, and this decision is randomly selected to be implemented, what will happen? (Assume you are one of the winners who will receive the remainder of the bonus payment from the randomly chosen decision.)”. The correct option is option 3.

In a few minutes you will be asked to go through the shopping platform, Woods. There are **two different supermarket layout options available** to you. In the next few pages, we will present those two options. For now, we just call them **option A and option B**.

On the next few pages, we present the two options and explain how to choose one of the options. Only then will you make your decisions.

(a) MPL Introduction - Seen First

For a better understanding, we will now show you an example. The table in which you will enter your real decisions will look like the one below.

You cannot select any options in this table. It is just an example.

Choose option A, and keep £0	☐	☐	Choose option B, and keep £2.00
Choose option A, and keep £0	☐	☐	Choose option B, and keep £1.00
Choose option A, and keep £0	☐	☐	Choose option B, and keep £0.00

How decisions work:

- Each row is a decision.
- For each decision, you will choose either Option A or Option B by ticking the left or right button on each row.
- Any decision of the 18 can be randomly selected to be applied, so consider each one carefully and give your true preferences.

Before continuing:

Please answer the comprehension question on the next page. You cannot move on until you answer it correctly.

(c) MPL Example (no context given for MPL) - Seen Third

You can choose between the two options as follows:

You will make 18 choices, each between two options.

In each choice, you pick either Option A or Option B. The only difference between the 18 choices is how much bonus money you would receive for each option.

Each question will ask: Would you rather choose Option A and keep a certain bonus amount, or Option B and keep a different bonus amount?

After making all 18 decisions, one will be randomly selected and applied. That means:

- For everyone, you will receive the shopping layout of that randomly chosen decision.
- If you are randomly selected as a winner, you will also receive the bonus from the randomly chosen decision (up to £7.00). In addition to your participation fee and one of your grocery baskets (from this week or last week) for real.

Since each choice could be the one that counts, please consider each decision carefully. We will now show you a practice example – this will not affect your real choices or outcomes.

You only make your real decisions after the next comprehension example.

(b) MPL Incentives - Seen Second

Comprehension check: If you choose Option B (the right-hand option) in the example decision below, and this decision is randomly selected to be implemented, what will happen? (Assume you are one of the winners who will receive the remainder of the bonus payment from the randomly chosen decision.)

Choose option A, and keep £0	☐	☐	Choose option B, and keep £2.00
------------------------------	-----------------------	-----------------------	---------------------------------

☐ In addition to your guaranteed participation fee, you will receive a **bonus payment of £0.00** and have the **shopping layout of option A**.

☐ In addition to your guaranteed participation fee, you will receive a **bonus payment of £2.00** and have the **shopping layout of option A**.

☐ In addition to your guaranteed participation fee, you will receive a **bonus payment of £2.00** and have the **shopping layout of option B**.

(d) MPL Comprehension Test (no context given for MPL) - Seen Last

Figure A.10: Screenshots of the MPL instructions shown to participants.

You answered the comprehension question correctly. Please now make your real decisions by selecting your preferred option in each row of the table.

Option A: Standard positioning of groceries. If you choose this option, when you access the online Woods grocery shopping platform, you'll see food items in the standard order. You will also have prices, eco and nutritional labels.

Option B: Eco positioning of groceries. If you choose this option, when you access the online Woods grocery shopping platform, you'll see food items in the order of most eco-friendly to least eco-friendly. You will also have prices, eco and nutritional labels.

Regardless of the order and option chosen, all the exact same food products will be made available to you.

	Option A: Standard Ranking		Option B: Eco Ranking	
	1	2		
Choose option A, and keep €0	☐	☐	Choose option B, and keep €7.00	
Choose option A, and keep €0	☐	☐	Choose option B, and keep €5.00	
Choose option A, and keep €0	☐	☐	Choose option B, and keep €4.00	
Choose option A, and keep €0	☐	☐	Choose option B, and keep €3.00	
Choose option A, and keep €0	☐	☐	Choose option B, and keep €2.50	
Choose option A, and keep €0	☐	☐	Choose option B, and keep €2.00	
Choose option A, and keep €0	☐	☐	Choose option B, and keep €1.50	
Choose option A, and keep €0	☐	☐	Choose option B, and keep €1.00	
Choose option A, and keep €0	☐	☐	Choose option B, and keep €0.50	
Choose option A, and keep €0	☐	☐	Choose option B, and keep €0.00	
Choose option A, and keep €0.50	☐	☐	Choose option B, and keep €0.00	
Choose option A, and keep €1.00	☐	☐	Choose option B, and keep €0.00	
Choose option A, and keep €1.50	☐	☐	Choose option B, and keep €0.00	
Choose option A, and keep €2.00	☐	☐	Choose option B, and keep €0.00	
Choose option A, and keep €2.50	☐	☐	Choose option B, and keep €0.00	
Choose option A, and keep €3.00	☐	☐	Choose option B, and keep €0.00	
Choose option A, and keep €4.00	☐	☐	Choose option B, and keep €0.00	
Choose option A, and keep €5.00	☐	☐	Choose option B, and keep €0.00	
Choose option A, and keep €7.00	☐	☐	Choose option B, and keep €0.00	

Figure A.11: Image of Incentivised MPL

E.2 Willingness-to-pay variable construction

Appendix E.2.1 documents how we construct the willingness-to-pay (WTP) measure from the week-2 multiple price list (MPL) and defines the number-of-switches measure used in the analysis, with robustness checks, and data selection discussion included in Appendix E.4.

E.2.1 Willingness-to-pay construction from the MPL

We construct an individual willingness-to-pay (WTP) measure from participants' week-2 multiple price list (MPL) responses. The MPL contains 19 decision rows. In each row, participants choose between Option A (standard sorting) and Option B (eco sorting). The two options differ only in the amount of the £7 endowment the participant keeps, which provides incentive compatibility for the stated choices.³⁵

Premium scale. We translate each MPL row to an implied premium p for eco sorting (Option B) relative to standard sorting (Option A). We define $p > 0$ to mean that eco sorting is more costly (the participant keeps less of the endowment under Option B than under Option A), and $p < 0$ to mean that eco sorting is cheaper (the participant keeps more of the endowment by choosing Option B).³⁶ The set of premium values used in the MPL is

$$p \in \{ \underbrace{-7}_{\text{MPL row 1}}, -5, -4, -3, -2.5, -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2, 2.5, 3, 4, 5, \underbrace{7}_{\text{row 19}} \}.$$

The MPL is ordered so that p is increasing in the row number (moving down the list). Given the extreme values at the endpoints, participants with monotone preferences were expected to choose Option B in early rows (large negative p , where eco sorting is subsidised) and, for sufficiently large positive p , switch to Option A.

³⁵In addition, the elicited preference affects the platform layout shown for the participant's second-week grocery shop.

³⁶For example, if under a given row the participant keeps £2 under Option A and £0 under Option B, by choosing Option B, that translates as a £2 premium and $p = 2$.

WTP definition. Let r^* denote the first row (as the premium p increases down the MPL) in which a participant chooses Option A after having chosen Option B in earlier rows. We define the participant’s WTP as the premium in the *last* row in which they chose Option B—that is, the largest premium at which the participant still prefers eco sorting.

We treat corner cases as follows. If a participant always chooses Option B (never switches), we code $WTP_i \geq 7$. If a participant always chooses Option A (never switches), we code $WTP_i < -7$. In the main analysis, these cases are handled as described in Appendix E.3.

Interpretation. Higher values of WTP_i indicate a stronger preference for eco sorting. Negative values indicate that the participant prefers standard sorting even when eco-sorting is subsidised relative to standard sorting (i.e., without requiring a sacrifice of the £7 endowment).

Number of switches. We also construct a measure of response consistency: the number of switches between Options A and B across the 19 rows. Under standard assumptions for MPL-based elicitation with a single underlying cutoff, participants should switch at most once. In practice, we observe corner solutions (0 switches) and non-monotone preferences (multiple switches), which we summarize using this count and illustrate with stylized examples in Appendix E.3. We also discuss how we deal with these cases for the analysis in Appendix E.4.

E.3 Stylised examples of the willingness to pay measures

In table A.13, for each of the MPL rows, we include column reporting a stylised MPL response pattern from a participant. We indicate the implied WTP and the number of switches, as discussed in the previous Appendix E.2.1.

Table A.13: Stylised MPL response patterns and implied WTP

Row	Option A (Standard sorting)	Option B (Eco sorting)	Always A	Always B	Prefers Stand. Sorting	Prefers Eco Sorting	Multiple switching
1	Keep £0	Keep £7.00	A	B	B	B	B
2	Keep £0	Keep £5.00	A	B	B	B	B
3	Keep £0	Keep £4.00	A	B	B	B	B
4	Keep £0	Keep £3.00	A	B	B	B	A
5	Keep £0	Keep £2.50	A	B	B	B	A
6	Keep £0	Keep £2.00	A	B	A	B	A
7	Keep £0	Keep £1.50	A	B	A	B	B
8	Keep £0	Keep £1.00	A	B	A	B	B
9	Keep £0	Keep £0.50	A	B	A	B	A
10	Keep £0	Keep £0	A	B	A	B	A
11	Keep £0.50	Keep £0	A	B	A	B	A
12	Keep £1.00	Keep £0	A	B	A	B	A
13	Keep £1.50	Keep £0	A	B	A	A	A
14	Keep £2.00	Keep £0	A	B	A	A	A
15	Keep £2.50	Keep £0	A	B	A	A	A
16	Keep £3.00	Keep £0	A	B	A	A	A
17	Keep £4.00	Keep £0	A	B	A	A	A
18	Keep £5.00	Keep £0	A	B	A	A	A
19	Keep £7.00	Keep £0	A	B	A	A	A
WTP	Last premium before first switch to A:		< -7	≥ 7	$p_{\text{row } 5} = -2.50$	$p_{\text{row } 12} = 1.00$	$p_{\text{row } 3}$
#Switches	Count of preference changes:		0	0	1	1	2

Notes: These are stylised examples of different possible responses to the MPL, and how they are processed for the WTP measure and the count of switches.

E.4 Data Selection from Multiple Price List

Of the 875 individuals for whom we elicit Multiple Price List (MPL) responses, 779 (89%) exhibit a single switch in choices across the MPL. Table A.14 reports the distribution of the number of switching events between the standard-sorting and eco-sorting layouts across participants. The experimental interface did not impose a single switching point on participants. Multiple switching is a common problem with MPL elicitation, with no clear correct ex-post analysis solution (Andersen et al., 2006; Yu et al., 2021; Jack et al., 2022; Bruner, 2011), notably with newer thought suggesting that multiple switching is not necessarily a signal of poor quality answers (Yu et al., 2021).

For the main analysis of the identity cue on participants' willingness-to-pay (WTP) for the eco-sorting layout, we retain respondents with three or fewer MPL preference switches, yielding an analysis sample of 850 participants. This follows the view in Yu et al. (2021) that repeated switching need not always reflect irrational or unusable responses. For this main specification, WTP follows the last-switch rule, that is, the highest premium at which a participant still chooses eco sorting before switching to standard sorting. Table A.15 indicates that this sample restriction is not identifiably heterogeneous across treatment arms. At the MPL elicitation stage, only the cue-then-MPL treatment arm differed in procedure (Section 2.4) from the other two conditions, with the addition of the identity cue for which we draw causality on the WTP. We therefore have no reason or evidence to believe that repeated switching is itself treatment-induced.

For robustness of our estimations, we also provide the results of willingness-to-pay measurements based on the sample of those with 0 or 1 preference switch only, yielding 828 participants. Within that narrower sample, the WTP is mechanically identical under the last-switch and first-switch rules for single-switch respondents. Due to the relatively small sample-size difference between these samples, we expect

limited changes in estimation. If an individual started at the top of the MPL preferring option A and then later switched to option B, this is considered irrational, and WTP is not calculated. If a person always chose option A throughout the MPL, then we assign a $WTP = -£7$.

For the main results study of the grocery basket impacts, participants with 0 switches will also be excluded, in addition to the multiple switchers. Those that have no preference switch across the MPL (either always preferring the eco-sorting or standard-sorting layout) would not be randomly assigned to either of the sorting layouts conditional on the WTP in our experiment design. e.g. if option 2 in the MPL were always chosen, then the participant would inevitably get the option 2 layout for the supermarket study.

Table A.14: Distribution of MPL answer switches

Nb. of switches	Freq.	Percent	Cumul. Percent
0	49	5.60	5.60
1	779	89.03	94.63
2	4	0.46	95.09
3	18	2.06	97.14
4	3	0.34	97.49
5	5	0.57	98.06
7	2	0.23	98.29
8	2	0.23	98.51
9	4	0.46	98.97
10	1	0.11	99.09
11	4	0.46	99.54
12	1	0.11	99.66
17	1	0.11	99.77
18	2	0.23	100.00
Total	875	100.00	

Table A.15: Distribution of MPL answer switches by cue treatment

Number of switches	Cue then MPL	MPL then Cue	Control (No Cue)	Total
0	18	10	21	49
1	263	261	255	779
2	0	0	4	4
3	4	8	6	18
4	2	1	0	3
5	3	1	1	5
7	0	1	1	2
8	1	0	1	2
9	1	2	1	4
10	0	0	1	1
11	3	0	1	4
12	0	0	1	1
17	0	0	1	1
18	1	0	1	2
Total	296	284	295	875

F Worked examples: from MPL choices to IPW propensity scores

To make the construction of the design-known propensity score $\hat{p}(w_i)$ in equation (3) concrete, Table A.16 works through six stylised, preference-consistent MPL response patterns. Columns C1 and C6 are the two consistent-corner cases (always-A and always-B), and columns C2–C5 are single-switchers whose switching point moves progressively down the MPL, from a strong preference for standard sorting (C2) to a strong preference for eco sorting (C5).

For each illustrative participant, the table reports (i) the row-by-row choice between Option A (standard sorting) and Option B (eco sorting); (ii) the implied last-switch WTP, following the convention defined in Appendix E.2.1; (iii) the design-known propensity score $\hat{p}(w_i) = \sum_{r: B \text{ chosen}} \pi_r$, computed using the unequal pre-specified row-draw probabilities ($\pi_1 = \pi_{19} = 18/53$ at the corner rows, $\pi_r = 1/53$ at each of the seventeen interior rows); (iv) the resulting IPW weight $1/\hat{p}(w_i)$ if the participant is drawn into the eco-sorted layout and $1/[1-\hat{p}(w_i)]$ if drawn into the standard layout; and (v) whether the participant is retained in the IPW basket-emissions sample. The two corner cases (C1, C6) are dropped because their propensity score is degenerate ($\hat{p} \in \{0, 1\}$): the random row draw cannot move them off their preferred layout, so the layout assignment is non-random and overlap fails by construction. The single-switchers C2–C5 retain interior propensity scores ranging from 0.34 to 0.66, with weights bounded between roughly 1.5 and 2.9, and are kept in the IPW estimation sample.

Table A.16: Worked examples: preference-consistent MPL patterns and the implied IPW propensity score

Row	Option A (Standard)	Option B (Eco)	C1 Always A	C2 Sw. row 2	C3 Sw. row 6	C4 Sw. row 11	C5 Sw. row 19	C6 Always B
1	Keep £0	Keep £7.00	A	B	B	B	B	B
2	Keep £0	Keep £5.00	A	A	B	B	B	B
3	Keep £0	Keep £4.00	A	A	B	B	B	B
4	Keep £0	Keep £3.00	A	A	B	B	B	B
5	Keep £0	Keep £2.50	A	A	B	B	B	B
6	Keep £0	Keep £2.00	A	A	A	B	B	B
7	Keep £0	Keep £1.50	A	A	A	B	B	B
8	Keep £0	Keep £1.00	A	A	A	B	B	B
9	Keep £0	Keep £0.50	A	A	A	B	B	B
10	Keep £0	Keep £0	A	A	A	B	B	B
11	Keep £0.50	Keep £0	A	A	A	A	B	B
12	Keep £1.00	Keep £0	A	A	A	A	B	B
13	Keep £1.50	Keep £0	A	A	A	A	B	B
14	Keep £2.00	Keep £0	A	A	A	A	B	B
15	Keep £2.50	Keep £0	A	A	A	A	B	B
16	Keep £3.00	Keep £0	A	A	A	A	B	B
17	Keep £4.00	Keep £0	A	A	A	A	B	B
18	Keep £5.00	Keep £0	A	A	A	A	B	B
19	Keep £7.00	Keep £0	A	A	A	A	A	B
#Switches			0	1	1	1	1	0
Implied WTP (£)	last-switch rule		< -7	-7	-2.50	0	5	≥ 7
$\hat{p}(w_i)$	eq. (3)		0	0.340	0.415	0.509	0.660	1
IPW wt. (Eco)	$1/\hat{p}$		—	2.94	2.41	1.96	1.51	1.00
IPW wt. (Std.)	$1/(1-\hat{p})$		1.00	1.51	1.71	2.04	2.94	—
Kept in IPW sample?	trim if $\hat{p} \in \{0, 1\}$		No	Yes	Yes	Yes	Yes	No

Notes: Stylised single-switcher and corner participants. $\hat{p}(w_i) = \sum_{r: B \text{ chosen}} \pi_r$ with $\pi_1 = \pi_{19} = 18/53$ and $\pi_r = 1/53$ otherwise. Corner cases (C1, C6) are trimmed because $\hat{p} \in \{0, 1\}$ implies non-random layout assignment.

G Survey Questions and Coding

Below, the table provides the survey questions asked to participants in the different weeks of the experiment, as well as the coding rules for the variables constructed from these questions.

G.1 Week 1 - Pre-Shopping

These are the questions asked to participants in the first week's survey, before their first shopping task.

Table A.17: Survey Questions and Variable Coding

Variable	Survey Question	Coding / Construction Rule
Age	What is your age (in years)?	4-group categorical: 1 " ≤ 31 ", 2 "32-41", 3 "42-52", 4 "53+"
Dietary Restrictions	Do you have any strict dietary requirements? You never eat any animal products (strict vegan) (1) – You can only eat specifically prepared food products (Halal or Kosher) (2) – Other (allergies including lactose, gluten, etc) (3) – None of the above (4)	Binary variable: 1 if any dietary restrictions are reported, 0 otherwise.

Continued on next page

Table A.17 continued

Variable	Survey Question	Coding / Construction Rule
Gender	What is your gender? Male (1) Female (2) Non-binary (3) Prefer not to say (4)	Binary variable (Female): 1 if Female, 0 otherwise.
Education	What is the highest educational qualification you have achieved so far in your life? (if your education is not from the UK, please select the equivalent type of qualification that suits): No formal qualifications (1) GSCE/O-Level (2) A-Level/Higher/BTEC (3) Vocational/NVQ (4) University-level (5) Other (7)	Binary variable (University-level): 1 if University-level education, 0 otherwise.
Shop for Others	For how many people do you buy food typically on a weekly basis, including partners, children, family? (1 means you mostly shop just for yourself)	Binary variable: 1 if shop for others too (> 1) and 0 if shop just for self.
Children	How many children live in your household? (Dropdown menu none to 7)	Binary variable (No Children): 1 if no children, 0 otherwise.
Personal Income	Please indicate your personal income bracket, in £ and pre-tax: o Less than £10,000 (1) £10,000 - £20,000 (2) £20,000 - £30,000 (3) £30,000 - £40,000 (4) £40,000 - £50,000 (5) £50,000 or more (6)	3-group categorical variable: 1: $\leq$ £30k; 2: £30k–50k; 3: $\geq$ £50k

Continued on next page

Table A.17 continued

Variable	Survey Question	Coding / Construction Rule
Weekly Grocery Spend	On average, how much do you spend on groceries for a week's worth of food? (in £ - GB pounds): Less than £25 a week; £25 - £50 a week; £50 - £75 a week; £75 - £125 a week; £125 - £175 a week; £175 - £250 a week ; More than £250 a week	Left with same categorical groups.
Risk Aversion	Imagine that you have a £100 of savings. You have the opportunity to invest in a risky financial product with the following features: There is a chance of 1 in 2 (50%) that the investment is profitable. If the investment is profitable, you earn 2.5 times the amount that you have invested. If the investment is not profitable, you lose the amount that you have invested. How much of your saved £100 would you like to invest in this product? (numerical input from 0 to 100)	Binary variable (Risk Averse): 1 if the participant chose value inferior to the median of the cohort, 0 otherwise.

G.2 Week 1 - Post-Shopping

These are the questions asked to participants in the first week's survey, after their first shopping task.

Table A.18: Survey Questions and Variable Coding

Variable	Survey Question	Coding / Construction Rule
Consumer Satisfac- tion	How satisfied are you with your shopping experience on the online grocery store today? Please give a score from 0 (very unhappy) to 10 (very happy)	Left as 11-point scale.
Green Identity	Please indicate your point of agreement with the following four statements, from strongly disagree to strongly agree. 1) Being a sustainable consumer is an important aspect of my way of being ; 2) I am the type of person who cares about the environment and sustainability ; 3) I think of myself as a sustainable consumer.	All individually collected on a 1–11 scale, and summed up to a 3–33 scale.
Hunger	Please tell us how hungry you are, from 0 (Not at all hungry) to 100 (Extremely hungry).	3 category variable: 1 if 0-33, 2 if 34-66, 3 if 67-100
Location	Where are you right now, while taking this survey?: At work; At home (not working); At home (working); Travelling (e.g. commuting on a train); During leisure time (e.g. at a cafe, park, social event, with family/friends); Other (please specify)	Binary variable (At Home not working): 1 if At home (not working), 0 otherwise.

G.3 Week 2 - Pre-Shopping

Other than week 2 consent and the Multiple Price List (MPL) elicitation, there were no other survey questions before the week 2 shopping task.

G.4 Week 2 - Post-Shopping

These are the questions asked to participants in the second week's survey, after their second shopping task.

Table A.19: Survey Questions and Variable Coding

Variable		Survey Question	Coding / Construction Rule
Consumer Satisfac- tion		Idem to Week 1	Idem to Week 1
Consumer Priorities		From the list of attributes below, please rank your shopping priorities. With (1) being most important (6) least important to you. (the list: Price, Taste, Convenience, Nutritional Value, Environmental Impact, Calories)	Left as numerical for each attribute.
Green Identity		Idem to Week 1	Idem to Week 1
Hunger		Idem to Week 1	Idem to Week 1

Continued on next page

Table A.19 continued

Variable	Survey Question	Coding / Construction Rule
Location	Idem to Week 1	Idem to Week 1
Green self-efficacy	On a scale from 1 (strongly disagree) to 7 (strongly agree), how much do you agree with the following statements?: 1) If I tried to behave in a climate-friendly way, that would have a positive effect on climate change. 2) I can personally help to reduce climate change by changing my behaviour. 3) I intend to reduce my consumption of animal products (e.g. meat, fish, dairy, eggs) in the near future? 4) I intend to reduce my consumption of other high CO2 and other greenhouse gas foods in the near future? 5) I intend to reduce my use of higher carbon transport (e.g. planes, cars) for less polluting options (e.g. public transport, bicycles)	Binary variable (Low green efficacy): 1 if the sum of the 5 items is below the median of the sample, 0 otherwise.
Mechanism Awareness Check	Did you notice a message displayed on the screen based on your shopping last week?: Yes; Not Sure; No	Left as 3 category variable.

H Appendix: Attention and Mechanism Checks

H.1 MPL Attention Check

In this Appendix section, we report the attention and mechanism checks collected at the end of the week 2 survey. Overall, 98% of participants passed the attention check, and over 93% answered the MPL comprehension question correctly on the first attempt, as previously mentioned in the paper and MPL-specific Appendix.

H.2 Identity Cue Attention Check

Table A.20 reports the distribution of responses to the cue message mechanism check question³⁷ by treatment arm. Responses differ across arms ($\chi^2(10) = 178.80$, $p < 0.001$) with participants in the cue treatment arms ("Cue Then MPL" and "MPL Then Cue" groups) responding "Yes" at rates of 38–52%, and no-cue arms at rates of 0.8–3%. The manipulation was salient to treated participants.

Table A.20: Mechanism check responses by treatment arm

Response	CueThenMPL Eco	CueThenMPL Standard	MPLThenCue Eco	MPLThenCue Standard	MPL NoCue Ctrl Eco	MPL NoCue Ctrl Standard
No	34 (21.7%)	35 (25.2%)	30 (22.4%)	40 (26.7%)	79 (47.6%)	68 (52.7%)
Not Sure	41 (26.1%)	51 (36.7%)	43 (32.1%)	52 (34.7%)	82 (49.4%)	60 (46.5%)
Yes	82 (52.2%)	53 (38.1%)	61 (45.5%)	58 (38.7%)	5 (3.0%)	1 (0.8%)
N	157	139	134	150	166	129

Notes: Cell entries are frequencies with column percentages in parentheses. Sample: week-2 respondents with a non-missing mechanism check answer. Pearson $\chi^2(10) = 178.80$, $p < 0.001$.

Of those that reported noticing the cue message (by replying "Yes" or "Not Sure"), 535 participants provided text responses describing what they recalled of the message.³⁸ We hand-coded these free-text answers into three themes: "Can't remember", "Environment related", and "Other". We report their distribution by treatment arm in Table A.21. Note that the column percentages are computed over the

³⁷"Did you notice a message displayed on the screen based on your shopping last week?"

³⁸Participants were asked "what did the message say?" if they had indicated "Yes" or "Not Sure" to the previous cue message attention check question.

aformentioned 535 participants. We note the higher relative prevalence of the “Environment related” responses in the cue treatment arms (Cue Then MPL and MPL Then Cue) relative to the control (no Cue) arms, as expected. The “other” category includes a variety of responses, including mentions of the spend and shopping list items, as well a spurious answer about a maths project.

Table A.21: Mechanism check (free-text) themes by treatment arm

Theme	CueThenMPL Eco	CueThenMPL Standard	MPLThenCue Eco	MPLThenCue Standard	MPL NoCue Ctrl Eco	MPL NoCue Ctrl Standard
Can't remember	21 (17.9%)	30 (29.7%)	20 (21.1%)	32 (31.4%)	36 (49.3%)	34 (72.3%)
Environment related	63 (53.8%)	48 (47.5%)	45 (47.4%)	53 (52.0%)	3 (4.1%)	0 (0.0%)
Other	33 (28.2%)	23 (22.8%)	30 (31.6%)	17 (16.7%)	34 (46.6%)	13 (27.7%)
N	117	101	95	102	73	47

Notes: Cell entries are frequencies with column percentages in parentheses. Free-text mechanism check answers were hand-coded into three themes (Can't remember, Environment related, Other). Sample: week-2 respondents with a non-missing free-text answer. Pearson $\chi^2(10) = 111.10$, $p < 0.001$.

Chapter 6 | Discussion and Conclusion

This thesis opened on the claim that behavioural heterogeneity, the differences in behavioural choices across individuals in the same choice environments, should be central to the design and evaluation of behavioural policy. Accounting for this behavioural heterogeneity allows us to better understand the distributional, ethical, dynamic and generalisability consequences of policies, and to design interventions that are more effective, fair, and scalable.

The three papers presented in this thesis are case studies observing and evaluating these behavioural heterogeneity impacts from choice environments, highlighting their prevalence across domains and stakes (labour choices versus daily and weekly meals). They show how simple average effects policy analysis can be misleading, or just lack nuances if not rigorous in exploring the extra behavioural heterogeneities that should also be a concern to behavioural policy researchers. These papers set out some evaluation and design pathways a behavioural policy maker can follow while remaining aware of central aspects of ethical, methodological and distributional concerns that the policy entails.

6.1 Overview

This thesis shows, through three papers, the importance of evaluating and designing choice environments for policy through a behavioural lens that has behavioural heterogeneity as the object of policy evaluation rather than a secondary check reported after the main average estimate. I take the perspective that the policy researcher, whether local or national government or even at the corporate firm level, wants to know not only whether a policy works, but for whom it works, through which channel, for how long, and at whose cost. Persistence, ethics and generalisability are then not a separate checklist running alongside heterogeneity, but the consequences of it.

However, as discussed in the literature review and presented in the papers, this is not a simple task. The behavioural policy researcher must navigate a broad set of ethical, practical, and methodological challenges to design and evaluate interventions that are effective, scalable, and ethically sound, while also considering the larger threads of economic, social and environmental impacts of the policies. These are important because, while there may be a large set of policies that could be effective for the immediate problem, the subset that satisfies all of the relevant criteria will typically be much smaller, and possibly empty.¹ In the case of an empty set, the policy researcher will have to triage and weight the policy priorities.

In this final chapter, I reflect on the contributions and evidence provided by the three papers. I situate my work within the broader landscape of behavioural research, with notes on limitations and potential avenues for further exploration. The three papers cover two largely different fields of application, labour choices and food choices, as a demonstration of the breadth of reach and importance of behavioural policy in both low- and high-stakes decision making. The papers also cover a range of methods, from observational data analysis to field experiments, and a range of policy interventions, from the labour market (happenstance) choice environments as a backdrop, to the use of identity cues in online grocery shopping. Choice environments are defined here in the broadest sense, as the context in which decisions are made, and the way that this context influences behaviour (by design or not). The world in which we make decisions, from high-stakes choices (such as jobs) to low-stakes choices (such as dinner for tonight), is a choice environment.

¹A caveat, of course, is whether the policy researcher can ex-ante evaluate the attributes across these themes and threads.

With Paper 1 (Chapter 3). I contribute to the discussion of thinking beyond ATE by highlighting how an incidental choice environment, here the labour market, has distributional risks and why this should be considered consequential for policy evaluation. In particular, I evaluate the role of the freelance market within the broader labour market choice environment. This is not a gig-platform website design, but a backdrop of labour market conditions (demands, opportunities and regulations). In this sense, I argue that even the choice of the next job is bent by the choice environment features of the labour market conditions and the personal situation of the job seeker. For example, someone who has lost their job and has low savings and a high need for income may be more likely to turn to quick-to-sign-on-to freelancing as a source of income, in the same way that someone in a desperate situation will exhibit less search patience on a website platform. The paper is observational (using cross-sectional empirical analysis, with no claim of causality), and the “policy” is the market situation itself rather than an intervention designed by a policy maker; the choice environment here is an accumulation of regulation and market conditions. The paper evaluates the behavioural consequences of this incidental backdrop.

With Paper 2 (Chapter 4). I shift focus to a slightly more traditional view of policy, specifically a deliberately designed choice environment provided by an entity rather than a market, which is the setting for a cafeteria food environment policy intervention. With this paper I show that even deliberately designed choice environments can have weak and unevenly distributed effects. And furthermore, that in this case the policy cost-effectiveness is muted further by seemingly having reduced persistence in effects once the policy is terminated. Specifically, I study a policy designed to encourage uptake of vegetarian meals. Compared with the labour choice context, this is a relatively low-stakes decision environment for individuals, although arguably at the aggregate level the stakes are high in terms

of environmental impact. This paper also involves higher-stakes contributions from the policy maker, in this case the cafeterias financing a subsidy. The methods step up from observational to a (natural) field setting, and the policy maker is an identifiable entity rather than a set of market conditions. The policy maker has actively designed an intervention, which I evaluate against relevant goals.

With Paper 3 (Chapter 5). I present an experimental study of a two-stage behavioural intervention. With this paper, I show how a deliberately designed behavioural intervention can be evaluated not only on average effects but also on channels of effect, heterogeneity, and ethical and scalability considerations. Where Papers 1 and 2 evaluated choice environments that already existed, in Paper 3 I take on the policy maker's role. I deliberately design a policy with a specific primary goal (reducing grocery-shopping emissions) and hold myself to the themes and threads of the policy maker, namely the ethical, replicability and generalisability considerations developed in the Literature Review (Chapter 2). Although the behavioural interventions themselves are traditional nudge-style instruments, choice-set ordering and messaging (with their combination somewhat more novel as an approach), the scope of the paper pushes the wider implications of intervention design and evaluation, in particular the measurement of the channels through which the intervention operates, its heterogeneity across consumer types, and an ethics and SANS-style appraisal of how the result might scale.

Taken together, the three papers trace a path from observing how an undesigned labour market choice environment shapes vulnerable workers' decisions, to evaluating a deliberately designed nudge in a food environment, to designing and stress-testing a behavioural intervention from scratch with the policy maker's full set of ethical and methodological constraints in mind. The methods escalate from observational, to natural field, to experimental. The stakes vary from labour market decisions, to single meal choices, to weekly grocery shopping. The papers also

span two sources of behavioural heterogeneity, set out in the Introduction Chapter of this thesis. Paper 1, with the incidental market for freelancing, is more of a study of individual *circumstance*, where income, job security and gender drive the results, although we do find that cognitive aspects as caught by the Cognitive Reflection Test (CRT) are important predictors of interest in freelancing. Papers 2 and 3 lead more with the *predisposition* heterogeneity, highlighting the importance of the baseline preferences cutting across any effects of vegetarian meal subsidies (Paper 2). In Paper 3, search effort being an important mechanism of effect for the sorting intervention, and the identity cue stimulating interest in the eco-sorting layout, both show the heterogeneity from predisposition. However, Paper 3 adds that a behavioural policy such as the eco-sorting layout can have a desirable effect across both circumstance and predisposition, while accounting for distributional, ethical or persistence concerns.

Contributions of the Thesis. The contributions of the work in this thesis are threefold. Empirically, each paper provides evidence of behavioural heterogeneity that affects the evaluation of the (designed or incidental) policy. Notably, we find a positive association between vulnerable individuals and intentions to freelance in Paper 1. In Paper 2, we find behavioural heterogeneity for the effects of a subsidy introduction and withdrawal based on baseline vegetarian preferences. Paper 3 identifies several behavioural heterogeneities, including a negative association between search effort and effects of re-sorting items in an online shopping platform. Methodologically, I argue in favour of channel measurement, pre-registration and incentive-compatible experiments designed to draw out and evaluate behavioural heterogeneities that are defensible and replicable. For policy, I highlight the important effects of distributional, ethical and dynamic consequences of behavioural heterogeneity, and that they should be considered at both policy design stage as well as the stage where the policy or market condition is at play.

6.2 Themes and Threads of the Three Papers

The three papers in this thesis attempt to address the question of how to design and evaluate behavioural public policies and choice architecture interventions.

In the literature review (Chapter 2), I identified some key themes for evaluating behavioural policies, each a consequence of behavioural heterogeneity. Ethics concerns itself with who is affected, what problem is being solved, and at whose or which cost; replicability and scalability need to know whether the behavioural heterogeneities in one sample are reflective of, or diverging from, another; and persistence asks for whom, if anyone, the observed effects last. I now apply these themes to the three papers.

For Ethics. The FORGOOD framework is the main cornerstone for evaluating behavioural policy discussed in the Literature Review chapter (Lades and Delaney, 2022). Of its seven parts (Fairness, Openness, Respect, Goals, Opinions, Options, Delegation), I focus on the five that are most binding in the settings studied, leaving Opinions and Delegation aside as they speak less directly to the policies considered here.

On Fairness, the three papers pull in different directions. As background, in the context of food taxes there is a well-studied regressiveness, as the tax affects a larger proportion of low-income households' incomes, for which tax recycling is a proposed solution (Klenert et al., 2023; Funke et al., 2022). Considering the papers in this thesis, there is a clear line of distributional analysis. Papers 1 and 2 both reveal distributional inequalities, but pulling in different directions. Paper 1 highlights that allowing freelancing as a form of self-employment within the labour market choice architecture may be a mixed blessing for vulnerable workers by providing a quick source of income but may also trap them in precarious work. Paper 2 highlights that the subsidy for vegetarian meals is unequal

in its benefits, with only the already-committed vegetarians benefitting. Paper 3 studies distributional effects across several factors and finds that, once treated, the re-sorting of items from most to least environmentally friendly is effective across several socio-demographic groups (age, gender, education), environmental preferences (green identity), and shopping types (low and high search-effort shoppers), although we do find a moderating effect of search effort on the sorting nudge, as would be expected from the sorting-nudge literature.

On Openness, Paper 3 is the most interesting case, as Paper 2's subsidy policy is not a background intervention but an upfront pecuniary intervention. Paper 1 is tricky to speak to as the policy and choice environment for the individual is incidental, not designed by any policymaker. However, freelancing as a job type was allowed as a legal form of work under the self-employment umbrella by a collective of policy makers. One could argue that allowing this form of work is transparent, but the pitfalls of the job type, such as the lack of worker protections discussed in the paper, may not be a transparent outcome of the choice to freelance. In the case of Paper 3, the proposed policy is not to force people or coercively change their choice environment (a concern in cases where, for example, a default sorting is changed). Instead, a self-nudge is designed. The cue message prompts an identity-salience shift toward environmental intentions (which is simply ignored when not relevant to the individual), and then offers the two sorting options. Naturally, for the experimental purposes the sorting allocation is randomised, but that is not the policy design proposed. The policy design is to offer the sorting options and let the individual choose, with the hope that the identity cue shifts some individuals toward the more environmentally sorted option.

On Respect, the autonomy of the decision maker is similarly preserved in Paper 3. If a participant self-selects into the re-sorting, the behavioural intervention then

does the effect-lifting through the Type 1 (automatic) pathway, which is less respectful of the decision maker's autonomy. But in cases of self-awareness of self-control issues, individuals can be happy to have temptations removed (Toussaert, 2018). By the same logic, people aware of their search foibles are the ones opting into the eco-sorted choice environment. In Papers 1 and 2, the incidental and designed policies both preserve the individual's autonomy under the case of assumed self-control. The choice of freelancing or vegetarian meals is not enforced. However, in Paper 1 it is discussed how vulnerable cohorts will likely be 'pushed' into considering freelancing for lack of time or ability to find formal work.

On Goals, the question of whether the policy is doing genuine good is most pointed for Paper 1 and the freelancing market. While freelancing provides a way to reduce unemployment friction and offer quick job access, it can be a mixed blessing, with some individuals 'pushed' into freelancing by their circumstances, and then potentially trapped in it by the style of freelancing chosen. On a broader analysis this is not necessarily desirable for the social policy maker, as freelancing does not offer the same regulations and protections as traditional employment, and from a distributional perspective it is the more vulnerable individuals who are most likely to be pushed into this precarious form of work. For the vegetarian subsidy of Paper 2, we see that there seems to be little to no long-term effects of the subsidy once it is withdrawn, and that the subsidy has not caused a change of preference, only a change of choice under a different pricing regime. While the subsidy does increase the uptake of vegetarian meals, this is focused among the more vegetarian individuals, which may not be the target audience if the policy wants long lasting changes in behaviour, and a shift to a more homogeneous flexitarian diet across the population. Paper 3, with a behavioural lens used for the design of the policy, seems to be the most 'good' intervention. Its intervention achieves the desired effects with little behavioural heterogeneity across population

factors, and no negative spillovers on budget or time spent on the platform.

On Options, cost-effectiveness and whether the outcomes justify the means are the central considerations when comparing policies, including between i-frame and s-frame interventions. Paper 2 discusses not only the distributional concerns with the subsidy but also its cost-effectiveness compared to classic label nudges of similar effect size. At the level of subsidy studied, I find little effect on behaviour, and effect sizes of this magnitude can be replicated for a fraction of the cost via, e.g., labelling interventions. From the perspective of effect size, the subsidy at the level provided is neither cheap nor effective, and so alternatives that optimise for one of these criteria might be considered better. With regards to Options in Paper 3, the intervention tries to combine strong Type 1 effects via self-selected sorting with Type 2 choice driven by an identity cue, while the cost of both components remains relatively low as they require only a website pop-up and an additional sorting order.

Replicability, Generalisability and Credibility of Evidence. With regards to the replicability of the evidence, the three papers in this thesis use a range of methods, from observational to experimental, with their own strengths and weaknesses. Papers 2 and 3 are primary data collection exercises, so before the data was collected for analysis, a pre-registered report for each of them was uploaded detailing the data type, the data analysis plan and the main research questions. This serves as a first-level safeguard against analytical researcher flexibility and p-hacking, as well as a way of clarifying the research questions and analysis plan before seeing the data. For Paper 1, the data had already been collected and seen by my co-author.

Looking at generalisability, a behavioural policy researcher should consider whether the study sample reflects the target policy population, and not the wider population (List, 2020). We consider this for all three papers. Paper 1 is an ob-

servational study, which limits the ability to make causal claims, but provides a real-world overview of the labour market with a large and rich micro-level dataset of survey data from across 16 countries, with country-level fixed effects and, for variables such as risk aversion and income, values which could be sensitive to country culture were measured relative to the individual's country median. The survey participants were recruited and paid for their participation according to local norms, as set by the survey company, and the sample was designed to be representative of the working population in each country. With this in mind, the findings are likely generalisable to the broader working population in these countries, although there may be some limitations due to self-selection bias (notably if there is a correlation between responding to surveys for money and sensitivity to freelancing opportunities). There is a further selection insofar as only individuals not currently in self-employment are studied, but this is precisely the target policy population for the question of who is at risk of being pushed into freelancing.

For Papers 2 and 3, more of a SANS-style analysis is required (List, 2020), as there are causal claims and the samples are smaller. For Paper 2, the sample is a natural field experiment in a real-world cafeteria setting, so behaviour is observed in a naturalistic environment with no concern about experimenter demand effects. However, the sample is limited to a student population (which deviates from the general population in age, life context, and income), and the subsidy applies to meals that are already heavily subsidised, so the discount is of a smaller magnitude than it would be in a real-world cafeteria or food setting. Porting the result beyond food is harder still, as food choice carries a cultural element. With regards to scaling, Mattauch et al. (2022) describe a situation in which equilibrium demand would shift not only because of the intervention but also because of a shift in social norms (for vegetarian meals, in this case). This could make the effect larger than predicted, as the social norm and endogenous preferences

act as a positive feedback loop. However, the effect size here is small and the intervention only affects the social group of already committed vegetarians, so social-norm spillovers to other groups are unlikely. Paper 3 recruits participants from a UK-wide online panel, with quotas for age and gender, and ensures that the individuals are the primary shoppers for their household and have used online shopping platforms previously. This reinforces sample representativeness for the population the online policy would target. Real food items and their metadata are used, and the grocery and willingness-to-pay choices are incentive-compatible, so the design meets the conditions of a framed field experiment and narrows the gap between the heterogeneity observed here and that a deployed policy would meet.

Vis-à-vis credibility of the results, Paper 3 (Chapter 5) is the most robust in its design, with a pre-registered report and an experimental design that allows for testing the main effect and the channels of effect. The channel analysis defends any observed outcome effect as linked to the expected channels, and not just driven by some other confounding factor. Specifically, I show that the identity-cue intervention updates identity beliefs, while the sorting intervention is strongest among those with low search effort.

Long-Run and Dynamic Considerations of Behavioural Policy. In the Literature Review (Chapter 2), I discussed the importance of considering the long-run and dynamic effects of behavioural policies. Behavioural policies might have short-term effects that fade over time, long-term effects that differ from the short-term effects, or effects that do not persist beyond the intervention period. These dynamics have important implications for the design and evaluation of behavioural policies, and for the policy maker's decision on whether to implement one policy over another. A policy that has a large short-term effect but fades quickly might be less desirable than a policy with a smaller short-term effect that

persists over time.

Long-run and dynamic considerations are particularly relevant for Papers 2 and 3, both of which study deliberate interventions. Although we can note that for Paper 1 the pool of vulnerable cohorts, and indeed wider population, interested in freelancing is unlikely to be depleting with every new generation of workers entering the job market. As a consequence the problem of people being ‘pushed’ into freelancing is unlikely to be a temporary one.

Paper 2 evaluates a subsidy for vegetarian meals, and finds that the effect is small but stable while the subsidy is in place, and that demand for the subsidised good drops once the subsidy is withdrawn. Unlike most other persistence-of-effect studies, the intervention was in place for more than a year and behaviour was then observed for a further year after withdrawal, so there is no concern that individuals were not given enough time to form new habits. This suggests that the subsidy operates through a traditional demand channel rather than by shifting endogenous preferences. And as the subsidy is not large enough to overcome the fixed costs of non-vegetarians, the policy does not shift preferences among those who might be most exposed to an identity shift, though this is beyond the scope of the paper.

Paper 3 evaluates a two-stage intervention combining an identity cue with a sorting nudge. Without any intervention, we observe a steep demand elasticity for the eco sorting around £0, with a sharp drop in demand between a willingness-to-pay of $-\text{£}0.50$ and $+\text{£}0.50$. The identity cue is effective at updating identity beliefs and at increasing the demand for the eco sorting. The sorting itself, however, is likely the main driver of persistence. As previously discussed, the sorting in Paper 3 (Chapter 5) can be considered as a default. The identity cue is necessary to shift preferences away from the standard sorting, but once the sorting has been changed to eco, switching back carries an effort cost that will

only be paid if it is smaller than the displeasure cost of the eco sorting. I measure satisfaction across sorting conditions and find no statistical difference between the eco and the standard sorting, suggesting that the change to eco sorting is likely to persist. Because the sorting operates through the Type 1 (automatic) pathway, the effect of the eco sorting is also likely to last, as people's bounded search effort is part of human nature.

6.3 Limitations

The paper-specific limitations are discussed in the respective papers, but here I will discuss some of the broader limitations for the thesis as a whole.

WEIRD and national populations. Excluding Paper 1, the other two papers are set firmly in the UK with a UK-resident population. Paper 2 takes place in a UK university cafeteria, and Paper 3 recruits from a UK-wide online panel. Whereas this is appropriate for the research questions and any UK-based policy contexts considered, the findings cannot necessarily be extrapolated to other countries. Even within the WEIRD context, food choices are culturally driven, so differences between, for example, the UK, the US, or France might generate different results. Studying non-WEIRD populations would add a further layer of deviation from the empirical evidence found here. Although the work stays close to theoretical work and other empirical studies, the thesis therefore cannot directly speak to behavioural policy design in these other contexts. Local practicalities, adoption patterns and regulations may also be a hindrance to the policies studied.

Power and granularity of the sample and data. Although each paper's sample is fit for its main research question, the type and scale of the samples constrain what the thesis can say more broadly, notably with regard to channels

of effect and finer subgroup analysis. In Paper 2, the small effect of the subsidy combined with a single-site sample limits the power to detect heterogeneity beyond the coarse vegetarian vs. non-vegetarian cut. As discussed in the paper, the inference is weakened further by the small number of treatment and control cafeterias. We also do not have individual-level information with which to do further heterogeneity analysis. In Paper 3, the sample is sufficient for the main effect and the headline channels, but finer heterogeneity slices (e.g. three-way interactions between factors) are necessarily under-powered.

No deployed at-scale policy trial. The methods escalate from observational, to natural field, to experimental across the three papers, but none of the three studies is a live, deployed policy intervention. Paper 1 measures intentions to freelance rather than actual behaviour, and does not test a proposed policy intervention to dampen the higher rate of vulnerable workers considering freelancing. Paper 2's sample is restricted to students. Paper 3 in particular is a simulated online retailer rather than a live deployment with a real shopping platform, although it is designed to be incentive compatible. The measurement of willingness-to-pay for eco sorting, instead of the binary sorting choice more typical of real-world online sorting options, is also a naturalness limitation of the design, albeit one that serves the important purpose of evaluating behaviour along a spectrum of willingness-to-pay. For the Paper 3 intervention we could also expect some substantial reaction from producers to the eco-sorting layout should it be popular with customers. Unfortunately, the paper only discusses partial equilibrium, like most experiments, in this case customer-side demand.

Comparability and practicality of policies. Returning to earlier discussions of megastudies and of cost-effectiveness comparisons across policies, the studies here do not have easy like-for-like comparisons because of differing testing environments and populations, but also because the reference points to which I have

access are the published papers, the number of which may be biased, as discussed in the publication-bias and study-to-RCT effect-gap literature (DellaVigna and Linos, 2022). The papers and their policies are also not designed in close collaboration with real policy makers to discuss their credible deployment and whether they are aligned with other current policies; as previously discussed, even good policies with large effects can be lost to organisational inertia (DellaVigna et al., 2024).

6.4 Behavioural Policy Design and Evaluation

Conclusions

From this thesis on behavioural policy design and evaluation, we understand that this is inherently a multi-criteria exercise, stemming from the awareness of behavioural heterogeneity. The policy maker must weigh ethics (FORGOOD), credibility (replicability, pre-registration, channel measurement), generalisability (SANS), distributional effects, and long-run dynamics all at once (Lades and Delaney, 2022; List, 2020). No single study, pilot, or even full rollout can satisfy all of these perfectly, and, as flagged at the opening of this chapter, the set of policies that satisfy all of them simultaneously is likely small and possibly empty. The way forward is then to triage and decide which criteria are binding for the problem the policy is trying to tackle.

Design-stage implications. Several lessons emerge from this research. First, measuring channels, and not only outcomes, is what gives credibility to the results and to any subsequent scaling claims. For example, Paper 3, with its identity-belief update and the search-effort interaction with sorting, is an important contribution to the evidence base on this front. Second, distributional analysis is coming to the forefront of behavioural policy evaluation, both *ex ante* and *ex*

post deployment. The cases discussed in this thesis (vulnerable workers drawn to freelancing, the subsidy benefitting already-vegetarian individuals, and the socio-demographically agnostic sorting intervention) are important for the Fairness and Goals criteria of FORGOOD. Third, cost-effectiveness and welfare design are important for the Options criterion of FORGOOD.

Evaluation-stage implications. During evaluation of a behavioural policy, pre-registration and pre-analysis plans help avoid analytical flexibility. Naturalistic settings and field experiments, rather than lab experiments, improve the Naturalness leg of SANS. Incentive-compatible measurements, such as Paper 3's willingness-to-pay design and the natural field setting of Paper 2 (as opposed to the stated intentions in Paper 1), support stronger inferences about policy-relevant behaviour. Measuring the channels of effect, as in Paper 3, is crucial for credibility and for understanding how the policy works.

Continued evaluation under rollout. Behavioural policy evaluation should not stop at adoption. It is important to remember that, because not all themes and threads can be evaluated pre-pilot or even during piloting, full rollout continues to offer opportunities to study the effects of the policy. A full rollout has a larger recipient population, may be wider-reaching, and offers more heterogeneity. Its stakes are necessarily at least as credible as those of any pilot. The policy will also likely be in place for longer. And finally, there will likely be slight changes in intervention design at rollout compared with when it was piloted, which could lead to slightly different effects. This is, in effect, a takeaway from Paper 1 where the labour market choice architecture is already in play, and I simply observe a possible correlation between the freelancing choice and worker vulnerability. These observations lead to suggestions for possible policy extensions and changes. During full rollout, a policy maker can not only observe aggregate or individual-level data live, but also probe for more information, perhaps through repeated survey

rounds collecting information on individuals' situations and preferences, and how these change over time.

6.5 Closing Remarks

To conclude on how to design and evaluate behavioural policies and choice environments through the lens of behavioural heterogeneity, I think the three papers taken together suggest that no single study or rollout can credibly satisfy all of the key criteria deriving from behavioural heterogeneity the policy maker should consider, whether at the behavioural policy experiment, pilot, or full rollout stage. But the policy maker can weigh several of them when evaluating a policy on its own or in comparison with another.

Paper 1 is a reminder that the behavioural influence of choice environments is already in play whether anyone designed it to do so or not, and that choice environments are never neutral. The environment will necessarily bend some people's preferences one way or another, and what is more, that effect is not necessarily the same for everyone. Paper 2 shows that even a deliberately designed and well-intentioned intervention can end up having small effects and poor cost-effectiveness compared to other behavioural policies. It also highlights the threat of heterogeneous policy benefits. Paper 3 shows that cheap behavioural interventions, such as an identity cue and a re-sorting of items, can have meaningful effects on behaviour, which defends a role for behavioural instruments in policy design.

Another previously highlighted point revolves around practicality of adoption, alongside cost-effectiveness, ethics, and the wider economic, social and environmental threads that the policy maker should be considering (Weber, 2013). The policy maker should be concerned with the threat of organisational inertia, which can ultimately decide which interventions ever see daylight irrespective of the

above listed threads and the general effectiveness of the policy itself (DellaVigna et al., 2024). But I think the research stage should resist letting adoptability become the binding criterion too early. If we only study the interventions that weave into current structures to dampen the risks of organisational inertia, we risk letting down society and missing the more powerful policies for which we should strive, or at the very least, use as a benchmark of potential.

Taken together, I think the contribution of the thesis is as much methodologically focused as it is about specific policy evaluation through the lens of behavioural heterogeneity. By treating behavioural heterogeneity as the object of evaluation rather than as auxiliary to the average, and by pursuing its analysis through its mechanism channels, distributional effects, ethical impacts and long-run dynamics, the policy maker gains a more complete understanding of a policy's effects, credibility and value to the target population. The job of the behavioural policy maker, then, is not only to design for what gets adopted, but to keep the benchmark of best-case policies in view, and to help the target population do “the right thing willingly” (Weber, 2013).

Bibliography

- Abadie, A., Athey, S., Imbens, G. W., and Wooldridge, J. M. (2023). When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, 138(1):1–35.
- Al-Ubaydli, O., List, J. A., and Suskind, D. L. (2017). What can we learn from experiments? understanding the threats to the scalability of experimental results. *American Economic Review*, 107(5):282–286.
- Allcott, H. (2015). Site selection bias in program evaluation. *The Quarterly Journal of Economics*, 130(3):1117–1165.
- Allcott, H. and Rogers, T. (2014). The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *American Economic Review*, 104(10):3003–3037.
- An, R. (2013). Effectiveness of subsidies in promoting healthy food purchases and consumption: a review of field experiments. *Public Health Nutrition*, 16(7):1215–1228.
- Andreoni, J. (1995). Cooperation in public-goods experiments: kindness or confusion? *American Economic Review*, pages 891–904.
- Andrews, I. and Kasy, M. (2019). Identification of and correction for publication bias. *American Economic Review*, 109(8):2766–2794.
- Angrist, J. D. and Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist’s companion*. Princeton University Press.
- Arkhangelsky, D., Athey, S., Hirshberg, D. A., Imbens, G. W., and Wager, S. (2021). Synthetic difference-in-differences. *American Economic Review*, 111(12):4088–4118.
- Ashraf, N., Camerer, C. F., and Loewenstein, G. (2005). Adam smith, behavioral economist. *Journal of Economic Perspectives*, 19(3):131–145.
- Banerjee, S. and John, P. (2024). Nudge plus: incorporating reflection into behavioral public policy. *Behavioural Public Policy*, 8(1):69–84.
- Banerjee, S. and John, P. (2025). Behavioral public policy: past, present, & future. *Policy and Society*, page puaf012.
- Barberis, N., Huang, M., and Thaler, R. H. (2006). Individual preferences, monetary gambles, and stock market participation: A case for narrow framing. *American Economic Review*, 96(4):1069–1090.
- Baron, J. (2018). A brief history of evidence-based policy. *The ANNALS of the American Academy of Political and Social Science*, 678(1):40–50.
- Barton, A. and Grüne-Yanoff, T. (2015). From libertarian paternalism to nudging—and beyond. *Review of Philosophy and Psychology*, 6(3):341–359.

- Bénabou, R. and Tirole, J. (2006). Incentives and prosocial behavior. *American Economic Review*, 96(5):1652–1678.
- Bénabou, R. and Tirole, J. (2011). Identity, morals, and taboos: Beliefs as assets. *The Quarterly Journal of Economics*, 126(2):805–855.
- Benartzi, S., Beshears, J., Milkman, K. L., Sunstein, C. R., Thaler, R. H., Shankar, M., Tucker-Ray, W., Congdon, W. J., and Galing, S. (2017). Should governments invest more in nudging? *Psychological Science*, 28(8):1041–1055.
- Bernedo, M., Ferraro, P. J., and Price, M. (2014). The persistent impacts of norm-based messaging and their implications for water conservation. *Journal of Consumer Policy*, 37(3):437–452.
- Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics*, 119(1):249–275.
- Bordalo, P., Gennaioli, N., and Shleifer, A. (2020). Memory, attention, and choice. *The Quarterly Journal of Economics*, 135(3):1399–1442.
- Bovens, L. (2009). The ethics of nudge. In *Preference change: Approaches from philosophy, economics and psychology*, pages 207–219. Springer.
- Callaway, B. and Sant’Anna, P. H. (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230.
- Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436.
- Camerer, C. F., Dreber, A., and Johannesson, M. (2019). Replication and other practices for improving scientific quality in experimental economics. In *Handbook of research methods and applications in experimental economics*, pages 83–102. Edward Elgar Publishing.
- Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427.
- Cameron, A. C. and Miller, D. L. (2015). A practitioner’s guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372.
- Chambers, C. D. and Tzavella, L. (2022). The past, present and future of registered reports. *Nature Human Behaviour*, 6(1):29–42.
- Chater, N. and Loewenstein, G. (2023). The i-frame and the s-frame: How focusing on individual-level solutions has led behavioral public policy astray. *Behavioral and Brain Sciences*, 46:e147.

- Chetty, R., Looney, A., and Kroft, K. (2009). Salience and taxation: Theory and evidence. *American Economic Review*, 99(4):1145–1177.
- Clark, M., Springmann, M., Rayner, M., Scarborough, P., Hill, J., Tilman, D., Macdiarmid, J. I., Fanzo, J., Bandy, L., and Harrington, R. A. (2022). Estimating the environmental impacts of 57,000 food products. *Proceedings of the National Academy of Sciences*, 119(33):e2120584119.
- Clark, M. A., Domingo, N. G., Colgan, K., Thakrar, S. K., Tilman, D., Lynch, J., Azevedo, I. L., and Hill, J. D. (2020). Global food system emissions could preclude achieving the 1.5 and 2°C climate change targets. *Science*, 370(6517):705–708.
- Cohn, A., Fehr, E., and Goette, L. (2015). Fair wages and effort provision: Combining evidence from a choice experiment and a field experiment. *Management Science*, 61(8):1777–1794.
- Colin Cameron, A. and Miller, D. L. (2015). A practitioner’s guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372.
- Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, 349.
- Connolly, D. J., Loewenstein, G., and Chater, N. (2025). An s-frame agenda for behavioral public policy research. *Behavioural Public Policy*, 9(3):593–613.
- D’Acunto, F., Malmendier, U., and Weber, M. (2021). Gender roles produce divergent economic expectations. *Proceedings of the National Academy of Sciences*, 118(21):e2008534118.
- De Chaisemartin, C. and d’Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–2996.
- Deaton, A. and Cartwright, N. (2018). Reflections on randomized control trials. *Social Science & Medicine*, 210(C).
- DellaVigna, S. (2009). Psychology and economics: Evidence from the field. *Journal of Economic Literature*, 47(2):315–372.
- DellaVigna, S., Kim, W., and Linos, E. (2024). Bottlenecks for evidence adoption. *Journal of Political Economy*, 132(8):2748–2789.
- DellaVigna, S., Lindner, A., Reizer, B., and Schmieder, J. F. (2017). Reference-dependent job search: Evidence from hungary. *The Quarterly Journal of Economics*, 132(4):1969–2018.
- DellaVigna, S. and Linos, E. (2022). Rcts to scale: Comprehensive evidence from two nudge units. *Econometrica*, 90(1):81–116.

- DellaVigna, S. and Paserman, M. D. (2005). Job search and impatience. *Journal of Labor Economics*, 23(3):527–588.
- DellaVigna, S. and Pope, D. (2018). Predicting experimental results: who knows what? *Journal of Political Economy*, 126(6):2410–2456.
- Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B. A., and Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, 112(50):15343–15347.
- Duflo, E. (2020). Field experiments and the practice of policy. *American Economic Review*, 110(7):1952–1973.
- Englmaier, F., Roider, A., and Sunde, U. (2017). The role of communication of performance schemes: Evidence from a field experiment. *Management Science*, 63(12):4061–4080.
- Epperson, R. and Gerster, A. (2026). Willful ignorance and moral behavior. *Management Science*.
- Evans, D. S. and Leighton, L. S. (1989). The determinants of changes in US self-employment, 1968–1987. *Small Business Economics*, 1:111–119.
- Fehr, E., Goette, L., and Zehnder, C. (2009). A behavioral account of the labor market: The role of fairness concerns. *Annu. Rev. Econ.*, 1(1):355–384.
- Ferraro, P. J., Miranda, J. J., and Price, M. K. (2011). The persistence of treatment effects with norm-based policy instruments: evidence from a randomized environmental policy experiment. *American Economic Review*, 101(3):318–322.
- Funke, F., Mattauch, L., Bijgaart, I. v. d., Godfray, H. C. J., Hepburn, C., Klenert, D., Springmann, M., and Treich, N. (2022). Toward optimal meat pricing: Is it time to tax meat consumption? *Review of Environmental Economics and Policy*, 16(2):219–240.
- García-Muros, X., Markandya, A., Romero-Jordán, D., and González-Eguino, M. (2017). The distributional effects of carbon-based food taxes. *Journal of Cleaner Production*, 140:996–1006.
- Garnett, E. E., Balmford, A., Marteau, T. M., Pilling, M. A., and Sandbrook, C. (2021). Price of change: Does a small alteration to the price of meat and vegetarian options affect their sales? *Journal of Environmental Psychology*, 75:101589.
- Garnett, E. E., Balmford, A., Sandbrook, C., Pilling, M. A., and Marteau, T. M. (2019). Impact of increasing vegetarian availability on meal selection and sales in cafeterias. *Proceedings of the National Academy of Sciences*, 116(42):20923–20929.

- Gelman, A. and Carlin, J. (2014). Beyond power calculations: Assessing type s (sign) and type m (magnitude) errors. *Perspectives on Psychological Science*, 9(6):641–651.
- Gneezy, U., Meier, S., and Rey-Biel, P. (2011). When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives*, 25(4):191–210.
- Gneezy, U. and Rustichini, A. (2000). Pay enough or don't pay at all. *The Quarterly Journal of Economics*, 115(3):791–810.
- Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277.
- Haaland, I., Roth, C., and Wohlfart, J. (2023). Designing information provision experiments. *Journal of Economic Literature*, 61(1):3–40.
- Harrison, G. W. and List, J. A. (2004). Field experiments. *Journal of Economic Literature*, 42(4):1009–1055.
- Hausman, D. M. and Welch, B. (2010). Debate: to nudge or not to nudge. *Journal of Political Philosophy*, 18(1).
- Hertwig, R. (2017). When to consider boosting: some rules for policy-makers. *Behavioural Public Policy*, 1(2):143–161.
- Hossain, T. and List, J. A. (2012). The behavioralist visits the factory: Increasing productivity using simple framing manipulations. *Management Science*, 58(12):2151–2167.
- Huntington-Klein, N., Arenas, A., Beam, E., Bertoni, M., Bloem, J. R., Burli, P., Chen, N., Grieco, P., Ekpe, G., Pugatch, T., et al. (2021). The influence of hidden researcher decisions in applied microeconomics. *Economic Inquiry*, 59(3):944–960.
- Imai, T., Toussaert, S., Baillon, A., Dreber, A., Ertac, S., Johannesson, M., Neyse, L., and Villeval, M. C. (2025). Pre-registration and pre-analysis plans in experimental economics. Technical report, JSTOR.
- Innocenti, S., Clark, G. L., and McGill, S. (2023). Experience of financial challenges, retirement concerns, and planning: evidence from representative samples of workers in 16 countries. *Journal of Pension Economics & Finance*, pages 1–19.
- Innocenti, S. and Golin, M. (2022). Human capital investment and perceived automation risks: Evidence from 16 countries. *Journal of Economic Behavior & Organization*, 195:27–41.
- Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8):e124.

- Johnson, E. J., Shu, S. B., Dellaert, B. G., Fox, C., Goldstein, D. G., Häubl, G., Larrick, R. P., Payne, J. W., Peters, E., Schkade, D., et al. (2012). Beyond nudges: Tools of a choice architecture. *Marketing Letters*, 23(2):487–504.
- Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.
- Klenert, D., Funke, F., and Cai, M. (2023). Meat taxes in europe can be designed to avoid overburdening low-income consumers. *Nature Food*, 4(10):894–901.
- Lacetera, N., Macis, M., and Slonim, R. (2012). Will there be blood? incentives and displacement effects in pro-social behavior. *American Economic Journal: Economic Policy*, 4(1):186–223.
- Lades, L. K. and Delaney, L. (2022). Nudge FORGOOD. *Behavioural Public Policy*, 6(1):75–94.
- Laibson, D. and List, J. A. (2015). Principles of (behavioral) economics. *American Economic Review*, 105(5):385–390.
- Levitt, S. D. and List, J. A. (2009). Field experiments in economics: The past, the present, and the future. *European Economic Review*, 53(1):1–18.
- List, J. A. (2011). Why economists should conduct field experiments and 14 tips for pulling one off. *Journal of Economic Perspectives*, 25(3):3–16.
- List, J. A. (2020). Non est disputandum de generalizability? a glimpse into the external validity trial. Technical report, National Bureau of Economic Research.
- Logg, J. M. and Dorison, C. A. (2021). Pre-registration: Weighing costs and benefits for researchers. *Organizational Behavior and Human Decision Processes*, 167:18–27.
- Lohmann, P. M., Gsottbauer, E., Doherty, A., and Kontoleon, A. (2022). Do carbon footprint labels promote climatarian diets? Evidence from a large-scale field experiment. *Journal of Environmental Economics and Management*, 114:102693.
- Ludwig, J., Kling, J. R., and Mullainathan, S. (2011). Mechanism experiments and policy evaluations. *Journal of Economic Perspectives*, 25(3):17–38.
- MacDonald, R. and Giazitzoglu, A. (2019). Youth, enterprise and precarity: or, what is, and what is wrong with, the ‘gig economy’? *Journal of Sociology*, 55(4):724–740.
- MacKinnon, J. G. and Webb, M. D. (2018). The wild bootstrap for few (treated) clusters. *The Econometrics Journal*, 21(2):114–135.
- Malmendier, U. and Nagel, S. (2011). Depression babies: Do macroeconomic experiences affect risk taking? *The Quarterly Journal of Economics*, 126(1):373–416.

- Marteau, T. M., Mantzari, E., and Hollands, G. J. (2025). Do nudges need a regulatory push? comparing the effectiveness and implementation of exemplar nudge (size-based) and non-nudge (price-based) dietary interventions. *Social Science & Medicine*, 373:118004.
- Mattauch, L., Hepburn, C., Spuler, F., and Stern, N. (2022). The economics of climate change with endogenous preferences. *Resource and Energy Economics*, 69:101312.
- Mertens, S., Herberz, M., Hahnel, U. J., and Brosch, T. (2022). The effectiveness of nudging: A meta-analysis of choice architecture interventions across behavioral domains. *Proceedings of the National Academy of Sciences*, 119(1):e2107346118.
- Milkman, K. L., Gromet, D., Ho, H., Kay, J. S., Lee, T. W., Pandiloski, P., Park, Y., Rai, A., Bazerman, M., Beshears, J., et al. (2021). Megastudies improve the impact of applied behavioural science. *Nature*, 600(7889):478–483.
- Mill, J. S. (1859). *On Liberty*. Ryerson University.
- Mullainathan, S. and Shafir, E. (2013). Decision making and policy in contexts of poverty. *Behavioral Foundations of Public Policy*, 16:281–300.
- Mullen, E. and Monin, B. (2016). Consistency versus licensing effects of past moral behavior. *Annual Review of Psychology*, 67(1):363–385.
- Panzone, L. A., Ulph, A., Hilton, D., Gortemaker, I., and Tajudeen, I. A. (2021a). Sustainable by design: choice architecture and the carbon footprint of grocery shopping. *Journal of Public Policy & Marketing*, 40(4):463–486.
- Panzone, L. A., Ulph, A., Zizzo, D. J., Hilton, D., and Clear, A. (2021b). The impact of environmental recall and carbon taxation on the carbon footprint of supermarket shopping. *Journal of Environmental Economics and Management*, 109:102137.
- Pechey, R., Hollands, G. J., and Marteau, T. M. (2022a). Explaining the effect on food selection of altering availability: two experimental studies on the role of relative preferences. *BMC Public Health*, 22(1):868.
- Pechey, R., Reynolds, J. P., Cook, B., Marteau, T. M., and Jebb, S. A. (2022b). Acceptability of policies to reduce consumption of red and processed meat: A population-based survey experiment. *Journal of Environmental Psychology*, 81:101817.
- Perino, G., Panzone, L. A., and Swanson, T. (2014). Motivation crowding in real consumption decisions: Who is messing with my groceries? *Economic Inquiry*, 52(2):592–607.
- Pizzo, A., Bauer, J. M., and Reisch, L. A. (2024). What shapes sustainable food choices? A field experiment on the impact of a behaviorally informed

- intervention and a price variation on sustainable food choices. *Available at SSRN*.
- Pollicino, D., Zamzow, H., Shreedhar, G., Galizzi, M. M., Naci, H., and Freitag, P. (2025). Can social norms promote sustainable food consumption? A systematic review. *Social Influence*, 20(1):2497341.
- Pritchett, L. and Sandefur, J. (2014). Context matters for size: why external validity claims and development practice do not mix. *Journal of Globalization and Development*, 4(2):161–197.
- Roodman, D., Nielsen, M. Ø., MacKinnon, J. G., and Webb, M. D. (2019). Fast and wild: Bootstrap inference in stata using boottest. *The Stata Journal*, 19(1):4–60.
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3):638.
- Roth, J., Sant’Anna, P. H., Bilinski, A., and Poe, J. (2023). What’s trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of Econometrics*, 235(2):2218–2244.
- Säll, S. (2018). Environmental food taxes and inequalities: Simulation of a meat tax in Sweden. *Food Policy*, 74:147–153.
- Schimmelpfennig, R., Vogt, S., Ehret, S., and Efferson, C. (2021). Promotion of behavioural change for health in a heterogeneous population. *Bulletin of the World Health Organization*, 99(11):819.
- Schulze-Tilling, A. (2025). The effectiveness of carbon labels. Technical report, University of Bonn and University of Mannheim, Germany.
- Shafir, E. (2013). *The behavioral foundations of public policy*. Princeton University Press.
- Shah, A. K., Shafir, E., and Mullainathan, S. (2015). Scarcity frames value. *Psychological Science*, 26(4):402–412.
- Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., et al. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1(3):337–356.
- Simoës, N., Crespo, N., and Moreira, S. B. (2016). Individual determinants of self-employment entry: What do we really know? *Journal of Economic Surveys*, 30(4):783–806.
- Simon, H. A. et al. (1972). Theories of bounded rationality. *Decision and Organization*, 1(1):161–176.
- Smith, A. (1759). *The theory of moral sentiments*. Penguin (2010 edition).

- Stechemesser, A., Koch, N., Mark, E., Dilger, E., Klösel, P., Menicacci, L., Nachtigall, D., Pretis, F., Ritter, N., Schwarz, M., et al. (2024). Climate policies that achieved major emission reductions: Global evidence from two decades. *Science*, 385(6711):884–892.
- Sunstein, C. R. (2014). *Why nudge?: The politics of libertarian paternalism*. Yale University Press.
- Sunstein, C. R. (2015). The ethics of nudging. *Yale Journal on Regulation*, 32:413.
- Sunstein, C. R. (2017). Nudges that fail. *Behavioural Public Policy*, 1(1):4–25.
- Sunstein, C. R., Reisch, L. A., and Kaiser, M. (2019). Trusting nudges? lessons from an international survey. *Journal of European Public Policy*, 26(10):1417–1443.
- Thaler, R. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior & Organization*, 1(1):39–60.
- Thaler, R. H. (2016). Behavioral economics: Past, present, and future. *American Economic Review*, 106(7):1577–1600.
- Thaler, R. H. (2018). Nudge, not sludge. *Science*, 361(6401):431–431.
- Thaler, R. H. and Benartzi, S. (2004). Save more tomorrow™: Using behavioral economics to increase employee saving. *Journal of Political Economy*, 112(S1):S164–S187.
- Thaler, R. H. and Sunstein, C. R. (2021). *Nudge: The final edition*. Penguin.
- Thaler, R. H., Sunstein, C. R., and Balz, J. P. (2013). *Choice architecture*. Princeton University Press.
- Thunström, L., Gilbert, B., and Ritten, C. J. (2018). Nudges that hurt those already hurting—distributional and unintended effects of salience nudges. *Journal of Economic Behavior & Organization*, 153:267–282.
- Toussaert, S. (2018). Eliciting temptation and self-control through menu choices: A lab experiment. *Econometrica*, 86(3):859–889.
- Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, 185(4157):1124–1131.
- Weber, E. (2013). Doing the right thing willingly. *The Behavioral Foundations of Public Policy*, 380.
- Weber, E. U. and Johnson, E. J. (2009). Mindful judgment and decision making. *Annual Review of Psychology*, 60(1):53–85.

- Wood, A. J., Graham, M., Lehdonvirta, V., and Hjorth, I. (2019). Good gig, bad gig: autonomy and algorithmic control in the global gig economy. *Work, Employment and Society*, 33(1):56–75.
- Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, 13(1):75–98.
- Zizzo, D. J., Parravano, M., Nakamura, R., Forwood, S., and Suhrcke, M. (2021). The impact of taxation and signposting on diet: an online field study with breakfast cereals and soft drinks. *Experimental Economics*, 24(4):1294–1324.

